Discrimination in Education: Methodology, Theory, and Empirics of Teachers' Stereotypes, Prejudice, and Discriminatory Behavior

Wenz, Sebastian E.

gesis
Leibniz-Institut
für Sozialwissenschaften

Schriftenreihe

Band 26

Sebastian E. Wenz

Discrimination in Education

Methodology, Theory, and Empirics of Teachers' Stereotypes, Prejudice, and Discriminatory Behavior

Discrimination in Education

GESIS Series
published by GESIS – Leibniz Institute for the Social Sciences

Volume 26

Sebastian E. Wenz

Discrimination in Education.
Methodology, Theory, and Empirics of Teachers' Stereotypes, Prejudice, and Discriminatory Behavior

Die vorliegende Arbeit wurde an der Fakultät Sozial- und Wirtschaftswissenschaften der Otto-Friedrich-Universität Bamberg als Inauguraldissertation zur Erlangung des akademischen Grades „doctor rerum politicarum" (Dr. rer. pol.) angenommen.

Sebastian E. Wenz

Discrimination in Education

Methodology, Theory, and Empirics of Teachers' Stereotypes, Prejudice, and Discriminatory Behavior

GESIS – Leibniz Institute for the Social Sciences 2020

Bibliographical information of the German National Library (DNB)

The German National Library lists this publication in the German National Bibliography; detailed bibliographical data are available via https://www.dnb.de.

ISBN	978-3-86819-044-1 (print)
ISBN	978-3-86819-043-4 (eBook)
ISSN	1869-2869

Publisher, printing and distribution:

GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8, 50667 Köln, Tel.: 0221 / 476 94 - 0
publications@gesis.org
Printed in Germany

To Kerstin, Ida, and Lotte
among whom I *discriminated* by time of coexistence
to set an order of appearance

Contents

Figures

Tables

Acknowledgments

This book is the result of a long journey. Along the way, many of the people I met in my professional and private life played a role in its completion.

I am very grateful for the advice and feedback on my work from the members of my committee—Cornelia Kristen, Corinna Kleinert, and Sandra Buchholz. I would also like to thank Hans-Peter Blossfeld for his support and advice over many years.

Professional advice and feedback on my work I have also received from many colleagues and friends including but certainly not limited to Christoph Homuth, Clemens Kroneberg, Anne Landhäußer, Thomas Leopold, Tobias Linberg, Tim Müller, Merlin Schaeffer, Steffen Schindler, Andreas Schmitz, Thorsten Schneider, and Volker Stocké.

I would like to thank my colleagues at the *NEPS* and later at the *Leibniz Institute for Educational Trajectories*, but especially Tobias Linberg, Vanessa Obermeier, Frank Goßmann, and Kerstin Hoenig for being great colleagues and friends.

I would not have finished this dissertation had my colleagues at *GESIS Training* at the *GESIS – Leibniz Institute for the Social Sciences* not supported me and given me the opportunity to work on it. For this, I am very grateful to the whole team at *GESIS Training*, but especially to Sören Petermann, Reinhard Schunck, and Nora Müller who repeatedly granted me time and space to work on my dissertation and supported me throughout. I am also very grateful to Angelika Ruf and Loretta Langendörfer for their support and understanding. Thanks to Bettina Zacharias and Philip Jost Janßen from *GESIS Publications* and Stefan Jünger for help during the publication process.

Thanks to all my friends in and outside of Bamberg, especially Tobias Linberg and Christoph Homuth for the support, friendship, and care. And sorry, especially to my friends outside of Bamberg, that I often just couldn't be there.

I also thank my whole family, especially my parents and brothers, who have supported me always no matter what. Sorry that I couldn't be there as often as we all would have liked.

And then there are my wife Kerstin and my daughters Ida and Lotte. Thank you for your love, care, and support. I have no words.

1 Setting The Scene for Research on Discrimination in German Education

> The primary function of the sociologist is to search out the determinants and consequences of diverse forms of social behavior.
>
> *(Merton, 1949)*

In this introductory chapter I aim at briefly setting the scene for my study on discrimination in German education. I reason that such a study is needed, how discrimination can be understood, and why both scientists and lay public *do* care and *should* care. I aim at showing that education is of particular importance when it comes to discrimination, inequality, inequity, and fairness, and which questions on discrimination in education I deem most interesting. I then present some methodological premises of my study. Many thoughts and arguments in later chapters are built on these premises. Finally, I give a brief outlook on the single chapters of this dissertation.

1.1 What is Discrimination?

Before I discuss different definitions and conceptualizations of discrimination in chapter 2, the reader may use the following as a basic and general working definition of discrimination for this introductory chapter: Discrimination is the act of treating two otherwise identical individuals differently based on any attribute, behavior, or characteristic that allows to distinguish these individuals (see, e.g., Blank et al., 2004; Heckman, 1998; Pager & Shepherd, 2008; Quillian, 2006, for similar conceptualizations). This is essentially a summary of popular definitions of discrimination. However, I will show in chapter 2 that—even though it is more useful than many other definitions—it has some problems that necessitate adaptation. As for an alternative, but—as I shall argue in chapter 2—not necessarily equivalent wording, the reader may think of discrimination as the *individual-level causal effect* of any attribute, behavior, or characteristic of an individual on how this individual is treated by another person. Both wordings are to be understood in a counterfactual sense and focus on *differential treatment* that may arise from treating a particular individual either more *negatively* or

more *positively* than it would have been treated in light of a counterfactual attribute, behavior, or characteristic.

In chapter 2 I shall argue that this very basic and general definition of discrimination as a causal effect is a much more useful starting point than many alternative definitions of discrimination put forth in the literature. However, even this definition I will criticize and adapt. In any case, to be useful for empirical research, the researcher needs to specify *which* attribute, behavior, or characteristic supposedly causes differential treatment. The most prominent example of such a cause or source of discrimination in the English and American literature has been a person's race, closely followed by sex or gender, respectively (see, e.g., Colella et al., 2017, for a review of 100 years of research on discrimination in psychology). In fact, most theoretical studies on discrimination have touched upon both race and gender (e.g., Aigner & Cain, 1977; Arrow, 1973; Becker, 1957/1971; J. R. Feagin & Booher Feagin, 1986; Levin & Levin, 1982; Phelps, 1972). For recent reviews that focus on racial discrimination see Pager and Shepherd (2008), Charles and Guryan (2011).

Discrimination caused by a person's race is usually called *racial discrimination* but some use the broader term *racism* instead. Discrimination by virtue of a person's sex or gender, respectively, is usually called sex or gender discrimination, respectively. Over time, especially sociologists have come to prefer the term gender over the term sex to highlight social and cultural components in stereotypes, prejudice, and discrimination against women. I shall use both terms to underline social and cultural factors but also the biological factors that contribute to both actual and perceived differences between men and women. Because the German literature is more concerned with *ethnic discrimination* (German: "ethnische Diskriminierung") instead of racial discrimination (see, e.g., Diehl & Fick, 2016, for a recent review), I shall use the term ethnic discrimination to refer to the situation of different groups of immigrants in Germany. Interestingly, discrimination based on a person's social class background (e.g., Jackson, 2009), sometimes discussed under the broader concept of *classism* (Lott, 2002), has received less attention and, if so, very often merely as mediating or confounding process in discrimination based on race or ethnicity (e.g., Bertrand & Mullainathan, 2004; Blalock, 1967; Mickelson, 2003; Myrdal, 1944). That racial or ethnic discrimination might be driven by social or socioeconomic factors is nevertheless an important observation that will be discussed at several occasions in this dissertation.

1.2 Why Discrimination?

Social scientists study discrimination typically—if not always explicitly—for two different reasons: First, discrimination on the basis of characteristics such as sex or gender, social class, and ethnicity, is of interest in its own right, as it violates norms prevalent in contemporary societies such as norms of fairness or meritocratic principles (Marsh et al., 2003; Rawls, 1971). Therefore, discrimination is usually considered unjust and unfair and sometimes explicitly defined as unjust or unfair treatment (see, e.g., Dovidio et al., 2010; Holzer & Ludwig, 2003). Many forms of discrimination that are considered unjust or unfair are also illegal in most developed countries (e.g., Chopin & Germaine, 2016; Fredman, 2012). Understood and motivated as unjust or unfair treatment, discrimination is a societal outcome that needs to be explained. Put differently, discrimination may be the *explanandum* in a sociological explanation.

Secondly, discrimination may be be part of the *explanans*: Sociologists and economists very often motivate research on discrimination with inequalities between social groups, such as blacks and whites or men and women, in various outcomes, such as wages, housing, or college admissions. Key questions in this dominant strand of the literature are: How can inequality theoretically be explained by discrimination and to what extent is inequality between groups actually due to discrimination? Both classic (Aigner & Cain, 1977; Becker, 1957/1971; Myrdal, 1944; Phelps, 1972) and more recent contributions (Carneiro et al., 2005; Heckman, 1998; Mickelson, 2003) argue over these question drawing on methodological, conceptual, and theoretical arguments as well as—last but not least—empirical evidence.

The distinction between discrimination as explanandum and discrimination as part of the explanans in an explanation of inequality between groups is virtually never made explicit and only sometimes discussed implicitly or touched upon. However, I find it crucial for a full understanding of how discrimination should be defined, identified, and estimated. That and how it matters, I will show in chapters 2 and 3.

1.3 Why Education?

The answer to the question *Why Education?* might simply be this: "Education makes life better." (Hout, 2012, p. 394). In fact, in Germany just like virtually anywhere else in the world, education has repeatedly shown to be positively associated with many individual and societal outcomes that are usually deemed positive such as occupational status and social class destination (Blau & Duncan, 1967; Breen & Jonsson, 2005; Ishida et al., 1995; Jackson et al., 2005; Klein, 2011; Müller & Pollak, 2004; Sewell et

al., 1970; Sewell et al., 1969; Sewell & Hauser, 1975), wages, earnings, and income (Brand & Xie, 2010; Card, 1999; Harmon et al., 2003; Psacharopoulos & Patrinos, 2004), higher likelihood of employment and lower likelihood of unemployment (Ashenfelter & Ham, 1979; Blundell et al., 1999; Mincer, 1991; OECD, 2016a), better health and various health related behaviors (Brunello et al., 2013; Brunello et al., 2016; Conti et al., 2010; von dem Knesebeck et al., 2006), measures of subjective well being including happiness and life satisfaction (Dolan et al., 2008; Kahneman & Krueger, 2006; Yang, 2008) and various social returns such as reduced crime rates (Chiras & Crea, 2004), increased political participation and civic engagement (Dee, 2004a; Henderson & Chatfield, 2011; Mayer, 2011; Verba et al., 1995), increased pro-environmental behavior (Meyer, 2015), as well as various liberal attitudes including support of freedom, pluralism, and democracy (Dee, 2004a; Robinson et al., 1999; Verba et al., 1995), and lower levels of anti-immigrant attitudes and racial prejudice (Biernat & Crandall, 1999; Carvacho et al., 2013; Quillian, 1995; S. L. Schneider, 2008; Wagner & Zick, 1995).

All of these associations are demonstrably at least in part causal effects—some direct, some indirect—in the counterfactual sense: Had individuals or states invested in and, thus, acquired, more (less) education, they would have had ended up with more (less) income, better (worse) health, lower (higher) crime rates, more (less) democratic citizens, and so on. While I, in contrast to Hout (2012), would like to avoid a normative judgment, most people would probably agree that these findings indeed suggest that education makes life better.

1.4 What Would We Want to Know About Discrimination in Education?

With regard to inequality in German education, it is a well established fact that inequality of educational opportunity and inequality of educational outcomes along the lines of *social class* or *socioeconomic status* are comparatively large. International studies on educational achievement in terms of obtained degrees and certificates as well as competencies such as literacy or numeracy have shown repeatedly that social inequality in German education is relatively high compared to other countries in both elementary and secondary school (e.g, Bos, Tarelli, et al., 2012; Bos, Wendt, et al., 2012; Breen & Jonsson, 2005; OECD, 2016b; Wendt et al., 2016), notwithstanding a—not always statistically significant—decrease in inequality over time both with regard to degrees (Breen & Jonsson, 2005; Breen et al., 2009) and competencies (Bos, Tarelli, et al., 2012; Prenzel et al., 2013; Wendt et al., 2016). Effect sizes for social class differences between students from lower or working class families and those from upper or upper middle class families in math and reading competencies lie around $d = .8$ at

the end of elementary school (e.g., Bos, Tarelli, et al., 2012; Bos, Wendt, et al., 2012; Stanat et al., 2012; Wendt et al., 2016, and my own calculations in chapter 5). The lower competencies of students from low social class families lead to worse grades and track recommendations for lower secondary school tracks. However, even conditional on competencies and other relevant covariates, numerous studies find that teachers award worse grades and recommend or prefer lower tracks for students from lower class families (e.g., Bos, Tarelli, et al., 2012; Bos, Wendt, et al., 2012; Ditton, 2013; Ditton et al., 2005; Maaz et al., 2011; Maaz et al., 2010; T. Schneider, 2011; Wendt et al., 2016). Surprisingly, there are only very few quantitative empirical studies that explicitly theorize and investigate social class discrimination or classism in German education (e.g., T. Schneider, 2011).

Similarly, the immigrant-native achievement gap in German education is larger than in many other countries around the world with regard to various measures of achievement such as years of schooling and highest degrees obtained (Dustmann et al., 2012; Heath et al., 2008) or competencies in reading, math, and science (Bos, Tarelli, et al., 2012; Bos, Wendt, et al., 2012; Marks, 2005; OECD, 2016b; Schnepf, 2007; Wendt et al., 2016). Just like in other countries, in Germany, too, the immigrant-native achievement gap is partly due to socioeconomic differences between immigrants and natives and, thus, is reduced once measures of socioeconomic status (SES) or social class are controlled for (Dustmann et al., 2012; Kristen & Granato, 2007; Marks, 2005; OECD, 2016b). However, usually and in Germany in particular, the disadvantage of immigrants cannot be fully explained by these factors—in fact, Germany turns out to have a comparatively large if not the largest immigrant-native achievement gap in competencies net of SES (e.g., Dustmann et al., 2012; OECD, 2016b). Effect sizes for the achievement gap in various competencies vary depending on the operationalization of immigrant status: Students with two parents born abroad lag behind about half a standard deviation ($d = .5$), students where only one parent is born abroad lag behind about a quarter ($d = .25$) or some third of a standard deviation ($d \approx .3$) (e.g., Bos, Tarelli, et al., 2012; Bos, Wendt, et al., 2012; OECD, 2016b; Stanat et al., 2012; Wendt et al., 2016). Looking at specific groups of immigrants, it turns out that the largest group, the immigrants of Turkish origin, but also other groups of guest workers—e.g., from the former state of Yugoslavia, Italians, Portuguese, Spanish—perform rather badly in the German education system with regard to different indicators (e.g., Kristen, 2002; Kristen & Granato, 2007; Olczyk, 2016): Students of Turkish origin are not only lagging behind students without immigrant background—with effects sizes of about one standard deviation in competencies (Stanat et al., 2012; Walter, 2009)—they also perform worse than the second largest group, students from the former Soviet Union, by more

than half a standard deviation (Stanat et al., 2012; Walter, 2009). The lower competencies of immigrants in general and the different groups of immigrants in particular result in worse grades and worse track recommendations compared to their peers without immigrant background (Kristen, 2006b). Depending on the ethnic groups examined and control variables used, residual differences in grades and recommendations remain (e.g., Gresch, 2012; Kiss, 2013; Kristen, 2006b; also see the overview in Diehl & Fick, 2016). In consequence, children with immigrant background in general and those of Turkish origin in particular overproportionally end up in lower secondary tracks (Diefenbach, 2010; Kristen, 2002, 2003; Kristen & Dollmann, 2009). However, even though the question whether or not teachers discriminate by virtue of students' ethnicity has been investigated and it seems that discrimination plays only a minor role in explaining inequality between ethnic groups in German education, evidence remains largely inconclusive due to several limitations of previous studies (see Diehl & Fick, 2016, for a review).

Less pronounced than both ethnic and socioeconomic achievement gaps are the differences in test scores, grades, track recommendations, track placement, and educational achievement between boys and girls. The pattern in tests scores and grades is such that boys outperform girls in mathematics and girls outperform boys in reading (Bos, Tarelli, et al., 2012; Bos, Wendt, et al., 2012; Prenzel et al., 2013; Reiss et al., 2016; Stanat et al., 2012; Wendt et al., 2016). Effect sizes of mean differences are about .1 standard deviation at the end of elementary school for both subjects (Bos, Tarelli, et al., 2012; Bos, Wendt, et al., 2012). At later stages in their educational career, the advantage of girls in reading is found to be larger than the advantage of boys in mathematics (Prenzel et al., 2013; Reiss et al., 2016). The same studies find that, over all subjects, girls increasingly outperform boys in grades, track recommendations, track placement, and educational achievement. Some studies find that boys receive lower grades conditional on test scores and other relevant controls (e.g., Hochweber, 2010; Maaz et al., 2011), other studies do not find such an effect (e.g., Wendt et al., 2016). By and large, observational studies suggest that, if anything, discriminatory grading to the disadvantage of boys is rather small in effect size. Similarly, some studies find statistical significant disadvantages of boys remaining in teachers' track recommendations or track preferences (e.g., Arnold et al., 2007; Ditton et al., 2005), but others—typically more recent studies—find no such effect (e.g., Bos, Tarelli, et al., 2012; T. Schneider, 2011).

For both students' social class and students' immigrant background or ethnicity, there is no conclusive evidence about the role of discrimination by teachers. Also, only one study implemented an experimental design to investigate ethnic discrimina-

tion in education using a sample of teachers (Sprietsma, 2013). Furthermore, we do not know much about teachers' stereotypes and prejudice—that is, the major determinants of discrimination—towards different groups of students. If teachers' stereotypes about characteristics of different groups of students are correct on average and they do not hold negative prejudice towards these groups, discrimination that disadvantages certain groups of students is rather unlikely, indeed. However, what if teachers' stereotypes are biased to the disadvantage of some groups and what if it can be shown that teachers do have negative prejudice towards particular groups—but maybe not others?

On the backdrop of the prevalent belief that discrimination by virtue of a person's social class background and ethnicity is considered unjust and unfair, and on the backdrop of inequalities in German education along the lines of social class and ethnicity, the general research questions in this dissertation are, whether there is evidence for discrimination against ethnic minorities in general, and students with a Turkish background in particular, or students from families of lower social classes in German education and, if so, what are the underlying mechanisms?

As for the question at which point in time discrimination in education should be of greatest interest, it seems relevant to recall that there is convincing evidence that in Germany, as in virtually all other developed countries, the first transition—the one from elementary to secondary school—is the most important in determining later levels of educational achievement and, thus, educational inequality but also outcomes in later life (e.g., Breen & Jonsson, 2005; Breen et al., 2009; Erikson & Jonsson, 1996; Shavit & Blossfeld, 1993). While later transitions and corrections to initial track placement are relatively less important in the sense that they show less unequal transition patterns of different groups, they add to, that is, exacerbate, the overall level of inequality between groups in German education (Buchholz & Schier, 2015; Hillmert & Jacob, 2005, 2010).

1.5 Methodological Foundations

1.5.1 Methodological Individualism

This dissertation is based on the principles of methodological individualism as proposed, refined, and advocated by many economists, philosophers, and sociologists (see, e.g., Udehn, 2002, for a brief history of methodological individualism). In this dissertation I adhere to a weak form of methodological individualism, similar to the

positions taken by, among others, Popper (1945, 1957), Boudon (1986a, 1986b), or Coleman (1986).

My perspective is very similar to what has been called institutional individualism (Agassi, 1975) and structural individualism (Wippler, 1978), respectively. These terms were introduced to highlight the differences to strong forms of methodological individualism, as advocated, among others, by Homans (1967, 1970), Hummell and Opp (1968), and Elster (1982), including psychologism (Mill, 1843) and other individualistic methodologies (Hummell & Opp, 1968; Menger, 1883).

Therefore, the key methodological principle I follow is this: Social phenomena, including discrimination, *should* be explained in terms of individuals, their physical and psychic states, actions, interactions, social, institutional, structural, and physical environment (see Udehn, 2002, cell 1b in figure 2). While this position implies that, in principle, all social phenomena can and, eventually, should be explained in terms of individuals, it acknowledges that, in a particular analysis, it is often not feasible to reduce the situation actors find themselves in to motives and general laws of human nature (Popper, 1945).

The claim that this situation may only bee seen as endogenous to individual action or behavior and, thus, to forbid to accept this situation as exogenous, would inevitably lead to an infinite regress taking us back all the way to a "beginning of society" (Popper, 1945). I reject this claim and allow the social situation to be exogenous to individual action. This way, the social, institutional, structural, and physical environment determines individual action and behavior by enabling, incentivizing, and constraining it.

Especially relevant for a study on discrimination is also to note that methodological individualism does not imply that the consequences of individual action or behavior are intended. Actually, methodological individualists typically stress the unintended consequences of human action or behavior—so do I. Hence, social phenomena are typically, at least partly, unintended consequences of actions of individuals. Even more so, what individuals do might not necessarily be properly described as intentional action, but—at least sometimes—more appropriately as automatic, spontaneous, or unconscious behavior (Boudon, 1998, 2003; Esser, 2001, 2009; Kroneberg, 2010; Kroneberg & Kalter, 2012). Mainly social psychologists but also sociologists and, recently, even economists, have pointed to automatic, spontaneous, and implicit mechanisms that determine discriminatory treatment (Bertrand et al., 2005; Devine, 1989; J. Feagin & Eckberg, 1980; Fiske, 1993b, 1998, 2000; Greenwald & Banaji, 1995).

The demand for microfoundations is a normative claim. It states that social phenomena *should* be explained in terms of individuals. I think that, ultimately, this claim

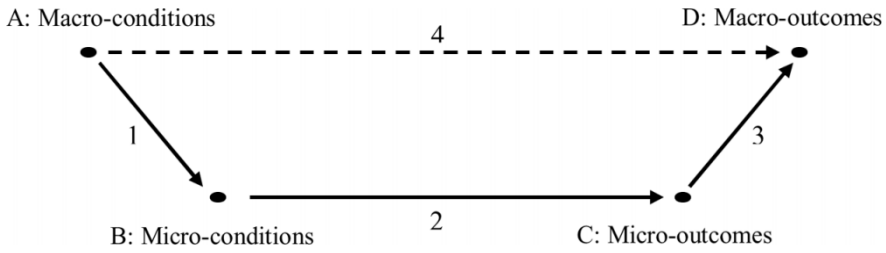


Figure 1.1: Representation of Coleman (1986)'s scheme from Raub et al. (2011).

is justified only insofar as microfoundations add anything to our understanding of the social phenomena we are studying. That is, it has to be shown that microfoundations make a difference. Following Udehn (2002, p. 501), this position coincides with viewing methodological individualism more as a "heuristic device or research program the fertility of which can only be ascertained a posteriori" than as an "a priori and universal principle".

However, in research on discrimination it is actually not difficult to show that microfoundations matter. In fact, assumptions or hypotheses about how individuals perceive, categorize, and, eventually, treat others based on the others' sex, social background, or ethnic background, matter a lot for both micro and macro outcomes. For those who are skeptical of methodological individualism as a universal principle in social science research, I will show—throughout this dissertation—why and how individuals matter in research on discrimination.

1.5.2 Model of Sociological Explanation

A schematic model for how to apply the rules of methodological individualism as outlined above, is the model of sociological explanation as advocated by Esser (1999). It builds on the macro-micro-macro scheme popularized by Coleman (1986) but already described by McClelland (1961) and others (see Raub et al., 2011, for a review of the scheme with focus on the links from macro to micro and from micro to macro).

According to the model, there are three major steps in every sociological explanation: First, the researcher has to investigate the "logic of the situation" (Popper, 1945) that applies to those individuals whose actions are to be explained. This involves a description of the situation (i.e., node A in figure 1.1), that is, the relevant social, institutional, structural, and physical environment. It also includes empirical or analytical bridge assumptions (arrow 1) about effects from the situation in A on the actors' knowledge, beliefs, attitudes, etc., and, thus, their more or less consciously perceived

set of alternatives (node B). In a study on discrimination in education, it has to be described in which situation teachers are, when they treat—and supposedly discriminate against—students of different background. The situation might be structured by laws or other rules that guide and constrain teachers in how to treat students, for example how to grade them or how to give track recommendations at the end of elementary school.

What follows is also known as "logic of selection" (Esser, 1999). Its key component is a micro-theory that explains how actors act or behave (arrow 2) under the given conditions. For instance, statistical discrimination theory (Aigner & Cain, 1977) would suggest that teachers judge and treat students according to a weighted sum of observed individual behavior and known group averages. Combining the logic of the situation and the logic of selection leads to predictions about behavioral outcomes (node C) that can be evaluated against empirical data. In a third step, the "logic of aggregation" (Esser, 1999) dictates that the behavior of individual teachers has to be aggregated to the macro-level (node D) following particular transformation rules (arrow 3). This step is considered very important but generally underrated (Coleman, 1986; Esser, 1999; Raub et al., 2011).

1.5.3 Value Judgments in the Study of Discrimination

I have already said that discrimination has been studied by many because it is considered unfair or unjust, or because it is illegal to discriminate against a person by virtue of characteristics such as race, ethnicity, or gender. Therefore, discrimination is a "value loaded term" (Myrdal, 1944, p. 214). Arrow (1998, p. 91) even goes so far as to claim that "[t]here is no way of separating completely the study of [...] discrimination [...] from moral feelings". Similarly, Quillian (2006, p. 300) notes that political ideology affects how discrimination is defined (see chapter 2).

However, following Hume (1738)'s dictum, there is no method, scientific or otherwise, to derive—without further assumptions—what ought (not) to be from what is (not). This holds for sociology as an empirical science (Weber, 1922) and, of course, it holds for a study on discrimination, too (Myrdal, 1944, p. 214). Actually, it strikes me that in a study on discrimination in particular, it is all the more "important to be analytic" (Arrow, 1998, p. 91) at all stages of the research process. I tried to be analytic, not political or moral, when I defined, identified, and estimated discrimination, and when I interpreted my empirical findings in this study. So, while it is nevertheless rather likely that my own moral feelings slipped in at some point, I hope that my arguments are convincing on scientific grounds.

1.6 How this Study is Structured

In chapter 2 I introduce and discuss various useful and some not so useful definitions of discrimination. I show how these definitions are related to the two distinct motivations for studying discrimination I have proposed in this introduction—discrimination as explanandum and as explanans. I show that understanding discrimination as a causal effect of an information about or a signal sent out by an individual on how this individual is treated by another individual is the most useful approach to the empirical study of discrimination. To this end, I make use of formal arguments from the recent literature on causality and causal inference about how to define and identify different causal effects.

In chapter 3, I review theories and models of discrimination from several disciplines including economics, social psychology, and sociology that might help understand why and predict whether teachers in German education discriminate among students by virtue of students' ethnicity, social class background, or gender. I discuss the general usefulness of the theories, existing evidence of whether actors actually behave according to the mechanisms suggested by the different approaches, and how they can be applied to the German education system.

In chapter 4, I discuss the central role of prejudice for understanding and predicting discriminatory behavior. I discuss the results and limitations of the few quantitative studies on explicit and implicit prejudice of teachers in German education towards different groups of students. Using one of these studies as a test case, I then present an analysis in which I quantify the bias in this study due to its geographically limited convenience sample of students. This is a limitation that, except one, all these studies have. To address this limitation and complement the findings of existing studies of teachers' prejudice, I show how to identify teachers and educators more generally in sufficient numbers in data from the German General Social Survey (GGSS/ALLBUS). I thereafter analyze teachers' prejudices towards different ethnic groups.

In chapter 5, I introduce an item battery to measure teachers' stereotypes about average competencies of different groups of students that I developed together with colleagues at the National Educational Panel Study (NEPS). I briefly discuss the role and functions of stereotypes in social cognition, intergroup relations, and, thus, discrimination in education and elsewhere. I then review in greater detail different conceptualizations of what stereotypes are and how they have been measured over time. Based on the definition we chose at the NEPS and I prefer in this study, I give a detailed account of the process of developing the new item battery. I present quantitative analyses that—based on theoretical considerations—speak to the validity of the

new instrument and allow to examine the accuracy of teachers' stereotypes towards different groups of students.

In chapter 6, I present results from analyses of experimental data that I collected in collaboration with Kerstin Hoenig and Anne Landhäußer to examine discrimination by teachers when assigning grades to essays and forming expectations about future performance of students signaling different ethnic background, social class background, and gender. I address several shortcomings of prior experimental research that all too often confounds social and ethnic discrimination by design, ignores the possibility of heterogeneous treatment effects across the distribution of ability, and is based on samples that heavily restrict the external validity of the findings.

I conclude in chapter 7.

2 Definitions of Discrimination

> The first thing to note is that discrimination is by no means easy to define concisely.
>
> *(Blalock, 1967)*

In this chapter I am concerned with questions of how to define and—to some extent—how to identify discrimination. My discussion will show why it is important to thoroughly think through what is meant by discrimination and to lay out definitions explicitly. In fact, many empirical studies on discrimination in general but also on discrimination in German education in particular seemingly fail with regard to the former and obviously fail with regard to the latter. At least in this regard, it seems, there has been only little, if any, change over the last decades, given Blalock (1967, p. 15) was right, when he wrote: "Many texts and descriptive works fail to attempt any definition at all".

2.1 On Useful and Not so Useful Definitions

My perspective on definitions and their role in the empirical social sciences is probably best explained in comparison to Popper (1945). I follow Popper (1945) in key aspects but do not agree without qualifications. In principle, I share Popper (1945)'s view that scientific definitions fundamentally differ from theories and hypotheses because they do not not make any empirical claims and, thus, can neither be true nor false. Also, definitions are not meant to grasp the essence of a term. I adopt Popper (1945)'s suggestion that scientific definitions are nominal definitions instead of essentialist definitions. In conclusion, I agree with Popper (1945) that the main purpose of scientific definitions is to provide "shorthand labels" to "cut a long story short".

However, one could argue that Popper (1945) would be skeptical of the exercise in this chapter, namely to ask which definition of discrimination we should adopt and which definitions we should not adopt in an empirical study on discrimination in German education. Such an endeavor might be seen as a violation of Popper (1945)'s principles, as it starts with the term discrimination, i.e., the definiendum, and seeks to find a definition, i.e., the definiens. Popper (1945) suggests that scientists should not and do not read a definition from left to right: Therefore, the question *What is discrimination?* "does not play any role in science" according to Popper (1945). Instead, scientific definitions are read from right to left—that is, they start with the definiens and pick

a definiendum as a short label. Thus, a relevant question—based, for the sake of an example, on the definition of discrimination from Levin and Levin (1982)—would be: *What should we call differential or unequal treatment of members of some group or category on the basis of their group membership rather than on the basis of their individual qualities?* The answer Levin and Levin (1982) gave, without asking the question, is discrimination. I would give the same answer, but I find the question to be ill posed. Put differently, I find their definition of discrimination—like many others—not very useful for empirical research in the social sciences.

So, maybe in contrast to Popper (1945), who suggests that "scientific or nominalist definitions do not contain any knowledge whatever, not even any 'opinion'", I think that definitions can be more or less *useful*. I say maybe, because Popper (1959/2004, pp. 15, 33–34) implicitly seem to share this perspective (also see Lakatos, 1980). Before I discuss various definitions of discrimination and why I find some of them more useful than others, here are my main criteria for evaluating how useful a definition of discrimination is. Probably the most important general criterion is that the definition should enable empirical researchers to answer their research questions. Therefore, a definition of discrimination should—amongst others—allow to test for different mechanisms of discrimination, to investigate discrimination against different groups, to examine the role discrimination plays in determining inequality, to assess the development of discrimination over time, and to compare discrimination across different contexts such as countries, federal states, schools, or neighborhoods. Many of the definitions I criticize and reject in the remainder of this chapter are not very useful because they do not help to answer these questions but make it difficult or even impossible to do so—some because they are too narrow, some because they are too broad, some for other reasons.

Also, I think that useful definitions should adhere to the methodological standards laid out in chapter 1. Most importantly, definitions of discrimination should not explicitly refer to or implicitly reflect any societal norms such as norms of fairness or meritocratic principles. Certainly, it is nevertheless legitimate that considerations of justice and fairness motivate research on discrimination.

Last but not least: While I think that the terms used by empirical social scientists do not need to match or reflect how they are used or understood by the lay public, it is—ceteris paribus—a good thing if we can reduce the costs of translating back and forth between scientific and public terminology.

2.2 Conceptualizing Discrimination: Premises

In this section, I lay out some premises on which my discussion of useful and not so useful definitions of discrimination is built. Many of these premises state which approaches I do *not* find useful in conceptualizing discrimination. I intend to get those less useful ideas out of the way before focusing in greater detail on more important and—not necessarily equivalent—more useful ideas.

2.2.1 Discrimination is About Behavior---Not About Attitudes Or Beliefs

Virtually every definition of discrimination refers to some form of behavior, action, or treatment. Or, as Pager and Shepherd (2008, p. 182) put it: "A key feature of any definition of discrimination is its focus on behavior." Therefore, I will, as is typically done and in line with the methodological principles discussed in chapter 1, assume that discrimination means that, eventually, an individual is doing something towards another individual. Note that this position is even shared in some contributions on so called institutional discrimination: "The "bottom line" in all types of discrimination is someone actually doing something to someone else" (J. R. Feagin & Booher Feagin, 1986, p. 25).

Since discrimination is about behavior, it is not equivalent with attitudes or beliefs and, thus, not equivalent with prejudice or stereotypes. Both analytically and empirically, sociologists and other social scientists have typically distinguished between these concepts. An early account of an empirical investigation is the classic study by LaPiere (1934) that shows that the relation of ethnic prejudice with ethnic discrimination may be very low. More recent meta-analyzes confirm that discrimination is only moderately correlated with both stereotypes and prejudice (Schütz & Six, 1996; Talaska et al., 2008). In the same vein, Merton (1949) argues that "[p]rejudicial attitudes not need [to] coincide with discriminatory behavior" (Merton, 1949, pp. 102–103) and presents a typology of ethnic prejudice and discrimination that includes the prejudiced non-discriminator as well as the non-prejudiced discriminator.

I suggest that a definition of discrimination shouldn't even refer to attitudes or beliefs. Defining discrimination as, for example, "the behavioral manifestation of prejudice" (J. M. Jones, 1997, p. 10) essentially rules out any other mechanism of discrimination. This would render meaningless any research on discrimination not based on prejudice, such as discrimination based on processes of stereotyping.

2.2.2 Discrimination is Not Necessarily Intentional

While especially earlier definitions conceptualized discrimination as *intentional* or *conscious* action (e.g., Aigner & Cain, 1977; Allport, 1954; Becker, 1957/1971; Blalock, 1967; Pincus, 1996) it is now widely agreed upon that this is a too narrow view on the empirical reality of social cognition, interpersonal behavior, and intergroup relations.

Based mainly on pioneering research by cognitive and social psychologists on processes of automatic, unconscious, implicit, or unintentional categorization (e.g., Allport, 1954; Devine, 1989; Fazio, 1990; Fiske et al., 1999; Greenwald & Banaji, 1995), today, social scientists from different fields agree that discrimination and its key determinants—stereotypes and prejudice—can be unconscious (Quillian, 2008), implicit (e.g., Bertrand et al., 2005; Greenwald & Krieger, 2006; Wittenbrink et al., 1997), automatic (e.g., Devine, 1989; Dovidio et al., 1997; Lepore & Brown, 1997), unintentional (e.g., J. Feagin & Eckberg, 1980), or subtle (e.g., Meertens & Pettigrew, 1997; Pettigrew & Meertens, 1995). For reviews on these forms of cognition, affect, and behavior see, e.g., Fazio and Olson (2003), Pager and Shepherd (2008), Quillian (2006).

Therefore, in contrast to Aigner and Cain (1977), Becker (1957/1971), Blalock (1967) and others, I do not limit the concept of discrimination to intentional or conscious behavior but treat unintentional or unconscious discrimination as equally discriminatory. In this study, for establishing discrimination, it does not matter whether a teacher intends to harm or consciously disadvantages a student. All that matters is whether and, if so, to which degree the student had been treated differently had they been of different ethnicity, class, or sex.

However, this is not to say that it cannot be interesting to distinguish between different forms of discrimination. Also, my position does not imply that intentional and unintentional acts of discrimination should be seen as morally equal. In fact, globally, a majority of people will probably not see them as morally equal, which might be reason enough for empirical researchers to investigate these forms separately. My position also does not mean that I reject theories or models that treat discrimination as intentional or conscious. In contrast, I will argue in this chapter and chapter 3 that, usually, such theories can easily be used to model both intentional and unintentional discrimination.

2.2.3 Discrimination is Not by Definition Unjust or Unfair

In chapter 1 I have argued that one of two major motivations to study discrimination is that in contemporary societies many consider discrimination based on variables such as sex, race, or class unfair and unjust. Therefore, it is not too surprising that

discrimination has also been defined as unjust or unfair treatment (see, e.g., Blank et al., 2004; Dovidio et al., 2010; Holzer & Ludwig, 2003; D. J. Schneider, 2004, for such conceptualizations).

However, above I have argued that definitions of discrimination should adhere to the methodological standards laid out in chapter 1 and should, thus, not refer to or reflect any societal norms or principles. I see two problems arising if scholars do so anyway: First, defining discrimination as unfair or unjust means to build a definition on normative and political terms. Since we have no scientific method to agree on what is just or fair and what is unjust or unfair, we are stuck with a problem that Simpson and Yinger (1972) summarized as follows:

> The essence of social discrimination is that there are some who say: we are "nicely distinguishing"; while others reply: no you are drawing "an unfair or injurious distinction" (Simpson & Yinger, 1972, p. 28)

Secondly, understanding discrimination as unjust or unfair, as something bad, something that should not be, something to reject and condemn probably explains why "some activists take all inequality among racial groups as discrimination" while "some conservative scholars, restrict discrimination only to acts that are intended to harm the target group" (Quillian, 2006, p. 300). Indeed, many definitions of discrimination are—obviously, apparently, or seemingly—build on the premise that discrimination is unjust or unfair. I intend to find a definition that is useful for empirical social science research and, therefore, build my discussion on a rather different premise, namely that discrimination is *not per se* unjust or unfair.

2.2.4 Discrimination is Not Inequality

We have already seen that this premise is less obvious than it might seem, but since "some activists take all inequality among racial groups as discrimination" (Quillian, 2006, p. 300), I feel the need to stress that, under any useful definition, discrimination is *not* the same as inequality. If it were, we wouldn't need a different term and questions on how discrimination and inequality are linked would all be meaningless. I shall return to the relation between discrimination and inequality below in section 2.4.1 when I discuss the distinction between individual discrimination and group discrimination. In chapter 3, I provide a more detailed discussion of how different theories of discrimination help to explain inequality between groups.

2.3 Discrimination as Causal Effect

That the question of whether or not discrimination of a particular kind exists, cannot be answered by a mere descriptive approach alone is no recent insight: "Definitions of discrimination usually, if not always, [...] require causal inferences" (Blalock, 1967, p. 15). But especially since the counterfactual or potential outcome model of causality became the standard approach to causality in the social sciences, more and more authors explicitly conceptualized discrimination in terms of causal effects. Blalock (1967)'s position is now widely shared in substantive contributions to the literature on discrimination (e.g., Blank et al., 2004; Heckman, 1998; Pager & Shepherd, 2008; Quillian, 2006) as well as methodological contributions to the literature on causality (e.g., Greiner & Rubin, 2010; Imai et al., 2013; Pearl, 2001, 2009; Pearl et al., 2016; D. B. Rubin, 1986; M. Sen & Wasow, 2016; VanderWeele & Hernán, 2012; Wang & Sobel, 2013). Some 40 years after Blalock (1967), Blank et al. (2004, p. 88), summarize: "Establishing that [...] discrimination did or did not occur requires causal inference".

2.3.1 Discrimination as Causal Effect: Foundations

The working definition of discrimination I gave in chapter 1—namely that discrimination is the individual-level causal effect of any attribute, behavior, or characteristic of an individual on how this individual is treated by another person—builds on various conceptualizations of discrimination as a causal effect (e.g., Blank et al., 2004; Heckman, 1998; Pager & Shepherd, 2008; Quillian, 2006). However, the definitions given by these and other authors differ at least slightly. To understand both differences and commonalities, I shall briefly recap the concepts of counterfactual causality and potential outcomes as well as the concepts of total, direct, and indirect effects before I discuss alternative conceptualizations of discrimination as a causal effect.

Individual-Level Causal Effects

The counterfactual or potential outcome framework is now the most widely accepted perspective on causality in the social sciences and beyond (Gangl, 2010; Imbens & Rubin, 2015; Morgan & Winship, 2015; Pearl, 2009; Pearl et al., 2016). The general idea is that a causal effect is defined as the difference in outcomes under a unit's factual state and one or more counterfactual states or, using potential outcome terminology, the difference between two or more potential outcomes under alternative causal states.

The individual-level causal effect or simply individual causal effect, δ_i[1], could then be written as

$$\delta_i \equiv y_i^1 - y_i^0, \tag{2.1}$$

where y_i^1 is the potential outcome of individual i in the treatment state, denoted by the right-hand superscript 1, and y_i^0 is the potential outcome of individual i in the control state, denoted by the right-hand superscript 0. The difference, δ_i, is the causal effect of treatment d_i, which is conceptualized as a variable that takes on at least two different values to potentially represent at least two alternative causal states—e.g., $d_i = 1$ if i is observed in the treatment group, and $d_i = 0$ if i is observed in the control group. Therefore, identifying and estimating a causal effect involves answering a—that is at least one—counterfactual question such as this one: What would have happened to individual i from the control (treatment) group, had individual i been in the treatment (control) group instead? The answer to this question is the total causal effect, or simply total effect, δ_i from equation 2.1, of the treatment, d, on the outcome, y. See section 2.3.1 below for more details on the distinction between total, direct, and indirect effects.

Population-level Causal Effects

If we take y_i^1, y_i^0, and d_i as individual realizations of population-level random variables Y^1, Y^0, and D, respectively, we can define the observable outcome variable Y as

$$Y = Y^1 \text{ if } D = 1,$$
$$Y = Y^0 \text{ if } D = 0.$$

This can be written as

$$Y = DY^1 + (1 - D)Y^0 \tag{2.2}$$

from which the biggest challenge for the counterfactual approach to causality becomes obvious: It is simply impossible to directly observe the effect of d on y, because it is logically impossible to observe one and the same individual or any other unit of interest in two or more different causal states at the same time. This "Fundamental Problem of Causal Inference" (Holland, 1986, p. 947) is the "fundamental reality of causal analysis" (Morgan & Winship, 2015, p. 45) and is typically addressed by defining and estimating some kind of *average causal effect* through aggregating over—usually,

1 Here, I mainly follow the notation from Morgan and Winship (2015). Elsewhere I also use notation from other authors.

but not necessarily—many individuals sampled from the population of interest.[2] The "broadest possible average effect" (Morgan & Winship, 2015, p. 46) is the *average treatment effect* (ATE) of D on Y:

$$ATE \equiv E[\delta_i] = E[\delta] = E[Y^1 - Y^0] = E[Y^1] - E[Y^0] \tag{2.3}$$

Here, the ATE stands for the average over all—possibly heterogeneous—individual-level causal total effects of D on Y in the population of interest.

Total, Direct, and Indirect Effects

In research on discrimination, the distinction between total, direct, and indirect effects is important and problematic at the same time. It is important because discrimination is often—though not always explicitly—defined as direct effect of, for example, race or gender on an outcome of interest such as wages or hiring decisions (e.g., Blank et al., 2004; Fix et al., 1993; Heckman, 1998; Quillian, 2006). Also, methodological contributions on the distinction between direct and indirect effects have used discrimination as an example of how to define, identify, and estimate direct and indirect effects (e.g., Imai et al., 2013; Pearl, 2001, 2009, 2014; VanderWeele & Hernán, 2012; Wang & Sobel, 2013). It is problematic since "the concepts of direct and indirect causal effects are generally ill-defined and often more deceptive than helpful" (D. B. Rubin, 2004, p. 162). It is the total effect that "is easiest to interpret, define and estimate" (Pearl, 2001, p. 411) and, thus, "[f]rom a counterfactual perspective, it is *only* the total effect of D on Y that has straightforward causal content" (Gangl, 2010, p. 28, my emphasis).

As said above in section 2.3, δ_i from equation 2.1 is the total effect of the treatment, d, on the outcome, y. In linear models with no interactions, the total effect, δ_i , represents the change in y caused by changing d by one unit.[3] That means that the total effect of d on y includes both the direct effect of d on y as well as all indirect effects that mediate the causal effect of d on y. Such a mediation is visualized in panel (a) of figure 2.1 in terms of population-level random variables D, Y, and M that represent treatment, outcome, and mediator, respectively. Panel (b) of figure 2.1 shows a slightly more complex mediation model with an additional mediator, N. Now, the

2 We might also estimate the individual causal effect or, more generally, unit causal effect, by observing the same individual or unit in different causal states over time.

3 When interactions are present or in the context of non-linear models, things are more complicated. However, a discussion of these issues is beyond the scope of this chapter. My arguments concerning the conceptualization of discrimination as causal effect are not affected by keeping things as simple as I do here.

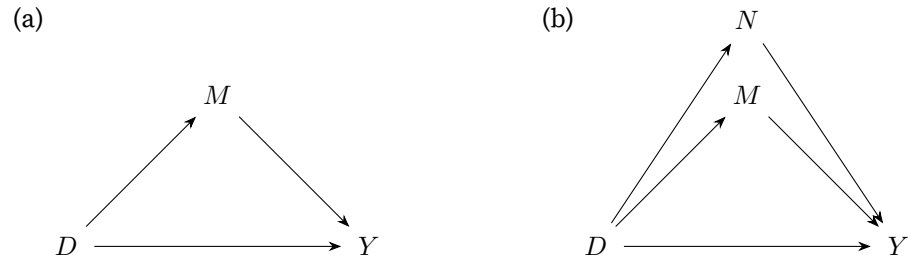


Figure 2.1: Panel (a) shows a simple mediation model with a treatment, D, a mediator, M, and an outcome, Y. Panel (b) shows a slightly more complex mediation model with an additional mediator, N.

appeal of the total effect is this: Whether the model in panel (a) or the model in panel (b) is assumed to be the correct model does not alter the definition or meaning of the total effect. Whatever the mechanism(s) that mediate the total effect, in both scenarios it is simply the familiar difference between two or more potential outcomes under alternative causal states, $d \in D$, namely $E[Y^1] - E[Y^0]$. It is this effect—the total effect—that is typically assessed in a controlled experiment (Pearl, 2001, p. 411).

In both panels of figure 2.1, the direct effect of D on Y is represented by the arrow pointing from D to Y, $D \rightarrow Y$[4]. In linear models with no interactions, it is defined and measured simply by the change in Y that occurs when D is changed by one unit while holding constant all other variables in the model including all intermediate variables. Put differently, the direct effect is the effect of D on Y net of the effects via all mechanisms represented by intermediate variables. Because, especially in nonlinear models, things can get pretty complicated, different kinds of direct effects are distinguished in the literature: pure direct effect (Robins & Greenland, 1992; Wang & Sobel, 2013) or natural direct effect (Pearl, 2001), controlled direct effect (Pearl, 2001; Wang & Sobel, 2013), and total direct effect (Robins & Greenland, 1992; Wang & Sobel, 2013). While a discussion of how to identify and estimate these different effects is not feasible at this point, it is important to understand that all forms of direct effects have one thing in common, namely that their substantive content depends on all other variables in the model. Two features are of particular importance: First, replacing one mediator with another—e.g., M in panel (a) with M^*—changes the substantive meaning of the direct effect, $D \rightarrow Y$. Secondly, adding an intermediate or mediating vari-

4 Each panel in figure 2.1 shows a directed acyclic graph (DAG). For introductions and discussions of their role in defining and identifying causal effects in the social sciences and beyond, see, among others, Elwert (2013), Morgan and Winship (2015), Pearl (2009), Pearl et al. (2016), Rohrer (2018)

able to the model that represents a mechanism by which D changes Y has the same consequences—it alters the substantive meaning of the direct effect, $D \rightarrow Y$. Thus, in figure 2.1, $D \rightarrow Y$ differs between panel (a), where it is the effect of D on Y net of M, and panel (b), where it is the effect of D on Y net of both M and N.

The intuition behind indirect effects or mediated effects is that they represent those and only those effects of the cause, D, on the outcome, Y, that operate through intermediate variables, such as M or N in figure 2.1. In panel (a) of figure 2.1, the path $D \rightarrow M \rightarrow Y$ constitutes an indirect effect of D on Y via M since D affects M and M affects Y in turn (Wang & Sobel, 2013, p. 215). In panel (b) there are two indirect effects of D on Y, namely $D \rightarrow M \rightarrow Y$ and $D \rightarrow N \rightarrow Y$. Quite obviously, there is no particular limit to the number of mediators and, thus, to the indirect effects of D on Y . Also, mediators may affect one another in various ways beyond what is shown in figure 2.1 (VanderWeele, 2015, chapter 5). As with direct effects, different—but not exactly the same—types of indirect effects are mentioned in the literature to account for different environments including linear and nonlinear models: pure indirect effect (Robins & Greenland, 1992; Wang & Sobel, 2013) or natural indirect effect (Pearl, 2001), and total indirect effect (Robins & Greenland, 1992; Wang & Sobel, 2013).

2.3.2 Discrimination as Causal Effect of Race, Gender, and Other Attributes

A crucial question with regard to the definition and identification of causal effects in general and the conceptualization of discrimination as a causal effect in particular is what precisely are the alternative causal states that—through their difference—define the causal effect of interest?[5] One seemingly natural and, thus, popular choice in the context of discrimination is to define treatment and control as belonging to different ethnic, racial, or social groups, or, more generally, possessing versus not possessing an attribute or possessing different attributes (cf. Holland, 1986). For the special case of racial discrimination, Blank et al. (2004, p. 79), for example, say they are interested in "the difference between two outcomes: the outcome if the individual were black and the outcome if the individual were white." According to Quillian (2006, p. 302), the relevant counterfactual question in research on discrimination is the following: "What would the treatment of target group members have been if they had been dominant group members?"[6]

5 See, e.g., Morgan and Winship (2015, pp. 37–43) or Imbens and Rubin (2015, pp. 3–5) for a discussion on how important it is to precisely lay out the different causal states.

6 Other statements in Blank et al. (2004), Quillian (2006) suggest that both are probably not interested in answering the counterfactual questions cited here but more narrow questions

In a study on ethnic discrimination in German education, the alternative causal states, $d \in D$, could be defined as being members of different ethnic groups. For instance, the treatment state, $D = 1$, would indicate that a student belongs to the Turkish ethnic minority, while the control state, $D = 0$, would indicate that a student belongs to the German ethnic majority. With regard to sex discrimination, the alternative causal states could be defined as being a girl, $D = 0$, or boy, $D = 1$, respectively. However, questions or statements like these have been challenged for various reasons (see, e.g., Greiner & Rubin, 2010; M. Sen & Wasow, 2016, for reviews of these debates): First, it has been argued that attributes in general and characteristics such as race, sex, or immigrant background in particular are "immutable" (Sobel, 1998, p. 334; Greiner & Rubin, 2010) and, therefore, not manipulable by any intervention (Berk, 2004; Freedman, 2004; Holland, 1986, 2003). Secondly, questions or statements such as the above cited by Blank et al. (2004), Quillian (2006) are often read as dealing with the total effect of attributes that are—from an essentialist or biological point of view (Greiner & Rubin, 2010, p. 776; M. Sen & Wasow, 2016, p. 500)—assigned very early in an individual's life; sex, for example, can be viewed as being assigned at conception. However, the total causal effect of a treatment assigned at conception or birth, the critics argue, is typically not of interest to many social scientists, especially not to those examining discrimination (Gangl, 2010; M. Sen & Wasow, 2016; VanderWeele & Hernán, 2012). Thirdly, the questions and statements from above are criticized for being not precise enough with regard to the actual treatment and the timing of treatment assignment and, thus, imprecise with regard to the alternative causal states (Greiner & Rubin, 2010; D. B. Rubin, 1986). I shall discuss these and other issues as well as proposed solutions in the next sections.

No causation without manipulation?

Probably the most famous—to some maybe: infamous—slogan from the literature on causality, "no causation without manipluation", seem to have first appeared in D. B. Rubin (1975, p. 238). According to Holland (1986), who repeats the slogan in capital letters ("NO CAUSATION WITHOUT MANIPULATION", Holland, 1986, p. 959), it was coined by Rubin and himself to emphasize that not everything can be a cause. Holland

---

that would, supposedly, best be answered by some kind of direct effect of, e.g., race on the outcome of interest. However, both Blank et al. (2004), Quillian (2006) do not discuss the contradiction between these questions and the precise effects they supposedly are after. Also, the cited questions make for a good start in a discussion of different alternative causal states in research on discrimination.

(1986, p. 946) argues that "[f]or causal inference, it is critical that each unit be *potentially exposable* to any one of the causes". In an own section on the question "What can be a cause?" he puts it this way:

> [...] I take the position that causes are only those things that could, in principle, be treatments in experiments. The qualification 'in principle' is important because practical, ethical, and other considerations might make some experiments infeasible, that is, limit us to contemplating *hypothetical experiments*. (Holland, 1986, pp. 954–955)

For traditional conceptualizations of discrimination as a causal effect of attributes of people on the way they are treated (as, e.g., in Blank et al., 2004; Quillian, 2006), this position poses a serious problem, since Holland (1986, p. 955) argues that only activities, e.g., being coached by teacher, but not attributes, e.g., a student's sex or race, could be treatments in experiments, not even in principle. Holland (1986, p. 955) reasons that the units of causal analysis—e.g., individual students—cannot be exposed to attributes, since "the only way for an attribute to change its value is for the unit to change in some way and no longer be the same unit". Therefore, according to (Holland, 1986, p. 955), "the notion of *potential exposability* does not apply" to attributes, which, in turn, rules them out as causal variables. Thus, a definition of discrimination as the causal effect of, for example, sex or race on the way an individual is treated, is not feasible.

While Holland (1986) does not discuss the phenomenon of discrimination explicitly, Holland (1988, 2003) do. In a comment on Dempster (1988), Holland (1988) argues that discrimination can and should be conceptualized as causal *effect*, not as causal *mechanism*—which is what Dempster (1988) suggests. However, Holland (1988) changes the question *What is discrimination?* into *What is the effect of discrimination?*, thereby avoiding a definition altogether. Holland (2003), who says that a counterfactual question such as "What would your life have been had your race been different? is so far from comprehensible that it is easily viewed as a ridiculous question" (Holland, 2003, p. 9), attempts an explicit definition of discrimination:

> [...] discrimination [is] a "statistical interaction" between a (potential) difference in societies and racial categories of people (Holland, 2003, p. 12).

This definition—that has neither been explicitly picked up by researchers who are substantively interested in discrimination nor by those working in the field of causality—does not conceptualize discrimination as differential treatment or any other form of behavior, but as the causal effect of societal, institutional, or systemic variables on racial inequality. By doing so, it misses the "the bottom line in all types of discrimi-

nation", namely: "someone actually doing something to someone else" (J. R. Feagin & Booher Feagin, 1986, p. 25). Since I, too, have argued in section 2.2 that any useful definition of discrimination should refer to some form of behavior, I can only reject Holland (2003)'s conceptualization.

Causation without manipulation? You bet!

One way of solving the problem of how to define discrimination in terms of causal effects of attributes such as sex or race, is to entirely reject the notion that causation requires manipulation (Gangl, 2010; Pearl, 2009; Pearl et al., 2016). Gangl (2010, p. 38), for example, argues that "[w]hether nonmanipulable factors such as gender, race, or class affect life courses is a perfectly sensible counterfactual question to begin with". Pearl (2009, p. 361) goes even further and argues that "many good ideas have been stifled or dismissed from causal analysis" since—and it is obvious that he wants to say: *because*—Holland (1986) promoted the phrase "no causation without manipulation". Pearl (2009) continues:

> Surely we have causation without manipulation. The moon causes tides, race causes discrimination, and sex causes the secretion of certain hormones and not others. (Pearl, 2009, p. 361)

Pearl (2009) has no problem with immutable characteristics as causes, since his perspective on causal effects is built around—potentially unrealistic—counterfactuals instead of more or less realistic interventions or manipulations. In fact, repeatedly, he has used the example of discrimination in general and sex discrimination in particular to explain his position on how to define and identify total, direct, and indirect effects of sex on outcomes such as college admissions (Pearl, 2009) or hiring decisions (e.g., Pearl, 2001; Pearl et al., 2016). Therefore, it is all but surprising that Pearl (2009, p. 362) advocates a "long-overdue counter-slogan: "Causation without manipulation? You bet!""

Indeed, on first sight, it seems perfectly sensible and legitimate to ask a question such as the following: *What would have been the track recommendation at the end of elementary school for a girl, had she been a boy instead?* However, upon closer inspection, this question—or, at least, a particular and popular reading of it (cf. D. B. Rubin, 1986)—is probably not of interest to many social scientists (Gangl, 2010, p. 38) and certainly not to those investigating discrimination (VanderWeele & Hernán, 2012, p. 109). For research on discrimination, this and similar counterfactual questions or statements are problematic if read as asking for the total effect of a treatment (M. Sen & Wasow, 2016; VanderWeele & Hernán, 2012) that occurs rather early in life—e.g., at

conception (Greiner & Rubin, 2010). Why exactly this is so problematic, I discuss in greater detail in the next paragraph.

The importance of time in defining alternative causal states

That things are a little more complex than suggested by Holland (1986), but also than they appear in the discussion of discrimination by Pearl (2001, 2009) and Pearl et al. (2016) was emphasized very early in the debate by D. B. Rubin (1986) in his comment on Holland (1986). While D. B. Rubin (1986, p. 962) upholds the motto "no causation without manipulation" as a "critical guideline for clear thinking in empirical studies for causal effects", he allows attributes of units to be treatments in hypothetical experiments as long as units, treatments, and outcomes are clearly defined. Maybe D. B. Rubin (1986) is simply more imaginative than Holland (1986, 1988, 2003) when accepting the statement that a male's life would have been different, had he been born a female instead—"whether because of some hypothetical Y to X chromosome treatment at conception, or massive doses of hormones in utero that would lead to female morphology at birth, or an at birth sex-change operation, or so forth" (D. B. Rubin, 1986, p. 961)—as causally meaningful.[7] D. B. Rubin (1986) brings up the example of sex discrimination in payment to show that, typically, things are more complex: He argues that the causal effect of being male instead of being female

> [...] has many possible versions ranging from some hypothetical 'at conception X to Y chromosome treatment' to replacing an 'F' with an 'M' on a job application form. (D. B. Rubin, 1986, p. 962)

While this is—to the best of my knowledge—the first time that discrimination is explicitly defined as causal effect of a signal (cf. Greiner & Rubin, 2010; M. Sen & Wasow, 2016), D. B. Rubin (1986)'s concern is of more general nature: D. B. Rubin (1986) and, even more explicitly, Greiner and Rubin (2010, p. 777) as well as Imbens and Rubin (2015, p. 5) make the case for the importance of *timing of treatment assignment* that, as a primitive of causal inference, is crucial in defining causal effects.

Figure 2.2 presents a visualization of D. B. Rubin (1986)'s point that timing of treatment assignment matters in general and in research on discrimination in particular: The total causal effect of being male instead of being female on the starting wage, Y, differs substantially between the two possible treatments mentioned by D. B. Rubin (1986)—some hypothetical treatment that changes the sex-determining chromosomes

7 For the role of imagination in defining causal effects, also see Imbens and Rubin (2015, p. 4).

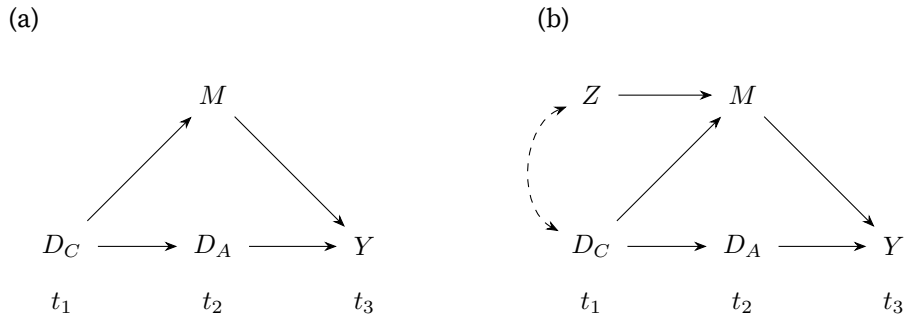


Figure 2.2: Building on D. B. Rubin (1986), the stylized DAG in panel (a) shows two versions of the causal effect of being male instead of being female—some hypothetical treatment that changes the sex-determining chromosomes from XX to XY at conception, D_C, versus changing a job application so that the applicant appears to be male instead of female, D_A—on starting wages, Y. The total causal effect of D_C on Y is made up of two indirect effects: $D_C \rightarrow D_A \rightarrow Y$ and $D_C \rightarrow M \rightarrow Y$, where M could be productivity. The total causal effect of D_A on Y is simply the direct effect $D_A \rightarrow Y$. The chronological order inherent to the DAG is highlighted by time points t_1, t_2, and t_3. Panel (b) shows a slightly more complex scenario with an additional covariate, Z, that shares an unobserved common cause with D_C and affects M and, thus, Y.

from XX to XY at conception, D_C, versus changing a job application so that the applicant appears to be male instead of female, D_A. The related counterfactual questions might then be the following: *What would have been the starting wage of a female employee, had she been a male all her life?* in case we were interested in the effect of D_C. In contrast, for the causal effect of D_A, the question might read *What would have been the starting wage of a female employee, had she appeared to be a male on the job application form?*

Now that the alternative causal states for the causal effects of both D_C and D_A are more clearly stated, the problem of the first question becomes more obvious (see panel (a) of figure 2.2): Under the plausible assumption that sex is randomly assigned at conception (Sobel, 1998, p. 335; VanderWeele & Hernán, 2012, p. 109; Gangl, 2010, p. 38), the effect of D_C on Y is not confounded by any variable such as parental education or social class. Because productivity, M, and the information about the applicant's sex on the job application form, D_A, are outcomes of D_C, we control for neither and the total causal effect of sex is defined and identified by the mere unadjusted wage difference between female and male employees. Certainly, this is neither equivalent to my understanding of discrimination nor to the understanding underlying the conceptualizations in most contributions in the literature. In contrast, the second ques-

tion, asking about the effect of D_A on Y, requires a different identification strategy, since there is an open backdoor path from D_A to Y, namely $D_A \leftarrow D_C \rightarrow M \rightarrow Y$ (see panel (a) of figure 2.2). Conditioning on M would block this backdoor path and, thus, identify the effect of D_A on Y. Intuitively, the different strategies make sense: The estimate of the effect of sex information from the job application on the wage is biased, if female and male job applicants actually have different productivity levels that also determine their wages. However, when interested in the effect of a sex change at conception, any resulting differences in productivity would be part of the effect of the sex change treatment and should not be held constant.

Whether truly "immutable" or not, the problem is virtually the same for most characteristics researchers in discrimination are typically interested in, such as race, ethnicity, immigrant background, but also measures of social class (background) and (parental) socioeconomic status. Since these variables are, in contrast to sex, not randomly assigned at conception, but are confounded with each other and further variables in determining productivity, the DAG in panel (b) of figure 2.2 is a more appropriate, still highly stylized, depiction of such a scenario. Let's say D_C is a person's race at conception, D_A is a racial signal on the job application—e.g., the applicants first name (Bertrand & Mullainathan, 2004)—, Y is the starting wage, M a person's productivity, and Z a measure of the person's social class background. To identify the total effect of D_C on Y in the scenario in panel (b), Z needs to be conditioned on. As in panel (a), the total effect of D_C on Y is mediated through D_A and M which is why conditioning on either of them would bias the estimate of the total effect and is, thus, prohibited. However, to identify the effect of D_A on Y, two backdoor paths need to be blocked: $D_A \leftarrow D_C \rightarrow M \rightarrow Y$ and $D_A \leftarrow D_C \dashleftarrow\dashrightarrow Z \rightarrow M \rightarrow Y$. Blocking both backdoor paths can be achieved either through conditioning on M or D_C.[8]

Discrimination in education: The role of time in defining causal effects

Applying the foregoing discussion, stimulated by D. B. Rubin (1986), to discrimination in education is straightforward: The question *What would have been the track recommendation at the end of elementary school for a girl, had she been a boy instead?* might have, at least, the following readings. First, it might ask for the causal effect of being conceived as a boy versus a girl. In this case, the question might be rephrased word-to-word and

8 Interestingly, conditioning on Z is not necessary in either case; note that, in case we condition on M only, we have to be sure that the only path from Z to Y is through M—if not, conditioning on Z or the intermediate variable between Z and Y would be necessary to identify the causal effect of D_A on Y.

read *What would have been the track recommendation at the end of elementary school for a female student, had she been conceived as a male instead?*—or, maybe even more vivid: *What would have been the track recommendation at the end of elementary school for a female student, had she been a male all her life instead?* Secondly, it might ask for the causal effect of being perceived as a boy versus a girl—through, for instance, some hypothetical and admittedly unrealistic perception changing experiment right before the teacher forms and gives the track recommendation. In this case, the general question might be rephrased to *What would have been the track recommendation at the end of elementary school for a female student, had she been perceived to be a male student by the teacher instead?*

Just like in the example on wage discrimination, the answer to the first question is simply the unadjusted difference in track recommendations between boys and girls: Being a boy instead of a girl since conception includes not only being born, but also being raised, educated, and socialized as a boy instead of as a girl, which, in turn, possibly results in different distributions of track recommendations by sex or, for that matter, gender. Put differently, conceptualizing discrimination as total causal effect of sex, as in this first question, means to equate discrimination with unconditional inequality. To answer the second question, a more sophisticated identification strategy would be needed: If it is true that boys and girls are raised, educated, and socialized differently by their parents, teachers, and society as a whole *and* that these differences—in, e.g., cognitive and noncognitive skills—affect the distribution of track recommendations, conditioning on them would be necessary to identify the causal effect of interest. In this scenario, discrimination is no longer equated with unconditional inequality but much closer to what is typically meant by discrimination in the literature (e.g., Blank et al., 2004; Quillian, 2006).

To sum up, simply rejecting the notion that causation requires manipulation and instead defining discrimination as the causal effect of attributes such as sex or race, does not strike me as a straightforward solution to the problem of how to conceptualize discrimination within a framework of counterfactual causality. For sex or gender but even more so for race, immigration background, or social class, simple counterfactual questions like "What would have happened to a nonwhite individual if he or she had been white?" (Blank et al., 2004, p. 77) are just not precise enough—mainly since they ignore the issue of timing of treatment assignment—and, thus, do not seem to be so "perfectly sensible" (Gangl, 2010) after all. Therefore, I will not adopt such definitions in this study.

Ceteris paribus terminology

Sometimes, some kind of *ceteris paribus* terminology is used in defining discrimination as causal effect. Take, for example, Quillian (2006, p. 302)'s definition of racial discrimination:

> Discrimination is the causal effect of race on an outcome with other factors held constant. (Quillian, 2006, p. 302)

Or, as another and even more prominent example, take Heckman (1998), who refers to the special cases of racial and gender discrimination:

> [...] discrimination is said to arise if an otherwise identical person is *treated differently by virtue of that person's race or gender*, and *race and gender by themselves have no direct effect on productivity*. Discrimination is a causal effect defined by a hypothetical *ceteris paribus* conceptual experiment—varying race but keeping all else constant. (Heckman, 1998, p. 102)

While the *ceteris paribus*-like phrases are probably meant to clarify things, they really don't. First, they do not solve the problem of carefully describing the alternative causal states. In fact, both definitions are perfectly compatible with sex or race being manipulated at conception or birth but also with the notion of manipulating signaled or perceived sex or race. Secondly, *ceteris paribus*-like phrases might have several meanings (Hausman, 1988): Adding *ceteris paribus* might mean to convey that SUTVA holds or that except for the treatment, other things (*ceteris*) are assumed to be equal (*paribus*) at the time of treatment assignment—but not, of course, thereafter (for discussions of the asymmetric nature of the ceteris paribus phrase see, e.g., Hausman, 1988, p. 313; Wooldridge, 2013, pp. 12, 205). In this case, the phrase is redundant, since this is precisely how individual total causal effects are defined. Alternatively, Quillian (2006) and Heckman (1998) may just want to stress that they are interested in the direct effect, not the total effect, of race and gender or sex. While this would be a rather vague way of doing so and might not be why Heckman (1998), Quillian (2006) use these phrases, I do indeed suggest that both are interested in conceptualizing discrimination as a direct effect. Before I discuss the problems of such conceptualizations below, I take a closer look at Heckman (1998)'s definition, since it features a noteworthy constraint: The effect of race or gender on productivity is assumed to be zero.

This is reflected in figure 2.3, where there is no arrow from D, representing race or gender, on P, representing productivity. This way, Heckman (1998) provides a very narrow definition of discrimination—more narrow are only those definitions that restrict discrimination to acts that intentionally harm individuals or groups (Quillian,

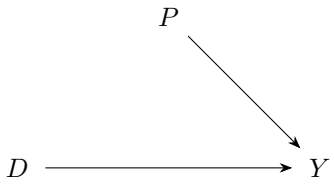


Figure 2.3: A DAG visualizing Heckman (1998, p. 102)'s definition of discrimination: Race or gender, D, directly affect how the individual's wage, Y, is set by the employer. Race and gender have no direct effect on productivity, P, that, of course, determines the wage.

2006, p. 300). Heckman (1998) probably—but not explicitly—seeks to rule out *statistical discrimination* (see, e.g., Aigner & Cain, 1977) and limit his definition to what, since Becker (1957/1971), is known as *taste discrimination* (also see Heckman & Siegelman, 1993, for a similar position). Heckman (1998) surely knows that there are direct effects of both race and gender on productivity, or—at the very least—confounding variables that induce an association of race and gender with productivity. Therefore, I read his definition as an analytical statement, not an empirical one. However, as a foundation for further efforts in identifying and estimating discrimination, such a strategy does not appear very useful to me and, thus, I will not adopt Heckman (1998)'s definition.

Discrimination as direct effect

One potential solution to the problem of conceptualizing discrimination as total effect of variables that are assigned very early in life—e.g., sex at conception—is to define discrimination as the *direct effect* of such variables. Based on the legal definition of discrimination in the US, Pearl (2001, 2009, 2014), Pearl et al. (2016), for instance, conceptualize discrimination explicitly as direct effect. Pearl et al. (2016) put it this way:

> Suppose, for example, we want to know whether and to what degree a company discriminates by gender (X) in its hiring practices (Y). Such discrimination would constitute a *direct effect of gender on hiring*, which is illegal in many cases. However, gender also affects hiring practices in other ways; often, for instance, women are more or less likely to go into a particular field than men, or to have achieved advanced degrees in that field. So gender may also have an indirect effect on hiring through the mediating variable of qualifications (Z). (Pearl et al., 2016, p. 76, my emphasis.)

Obviously, this definition suffers from the same imprecision in articulating the alter-

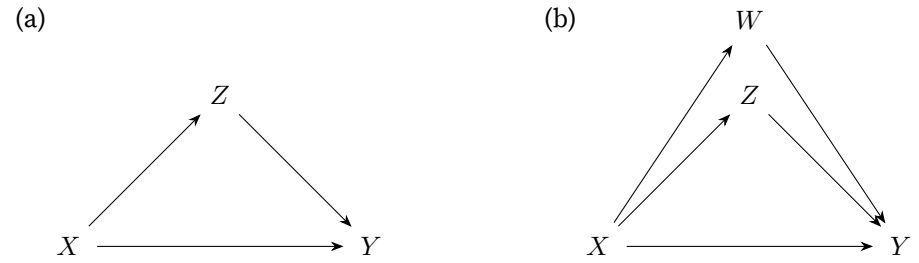


Figure 2.4: The DAG in panel (a) shows the model by Pearl et al. (2016): Gender, X, affects hiring, Y, directly but also indirectly through qualifications, Z. The direct effect, $X \rightarrow Y$, is what constitutes discrimination according to Pearl et al. (2016). The DAG in panel (b) shows a slightly more complex mediation model with an additional mediator, W. Adding W to the model changes the substantive content of the direct effect of gender on hiring, $X \rightarrow Y$, and, thus, the definition of discrimination as given by Pearl et al. (2016).

native causal states as the definitions in the section above: It is not clear what the treatment is supposed to be exactly and when it is meant to occur—maybe "some hypothetical 'at conception X to Y chromosome treatment'" (D. B. Rubin, 1986, p. 962)? But then sex would have been the more appropriate term than gender, wouldn't it? We are certainly not talking about "replacing an 'F' with an 'M' on a job application form" (D. B. Rubin, 1986, p. 962), since such a manipulation, obviously, does not affect the actual qualifications. Now, to be fair, Pearl et al. (2016) do not intend to contribute substantively to the literature on discrimination. I discuss their definition mainly because it is—in sharp contrast to the definitions in most substantive contributions—rather explicit about the direct effect conceptualization of discrimination.

In principle, figure 2.4 features the same setup as figure 2.1 above. The simple mediation model depicted in panel (a) of figure 2.4 shows the conceptualization by Pearl et al. (2016, figure 3.11): Gender, X, affects hiring, Y, directly but also indirectly through qualifications, Z. The direct effect, $X \rightarrow Y$, is what constitutes discrimination according to Pearl et al. (2016). However, such a definition is problematic, because the substantive meaning of direct effects depends on all other variables in the model. First, changing Z changes the substantive meaning of $X \rightarrow Y$ and, thus, the definition of discrimination. With regard to discrimination in hiring, Z might also be productivity. With regard to discrimination in track recommendations at the end of elementary school, X might be the social class background of the students, and Z could be educational achievement. But since the definition of Pearl et al. (2016) does not provide any general rule for choosing Z appropriately in different contexts, it might as well be argued to be something else.

Secondly, adding a mediating variable to the model also changes $X \rightarrow Y$. Let's say, W in panel (b) of figure 2.4 is job interview performance. Then, the direct effect of X on Y no longer captures but is conditional on the differences between men and women in job interview performance. This might actually be seen as a more appropriate definition of discrimination in hiring, but without a general definition of what is meant by discrimination, it is not clear how to argue in favor of changing the definition from $X \rightarrow Y$ in panel (a) to $X \rightarrow Y$ in panel (b). The same problem occurs in a study on social class discrimination in education. Say, we add to educational achievement, Z, parental support, W, then we face the same problem: We have no general definition of discrimination and, hence, no general rule to decide whether parental support should be hold constant or not. So, while we might accept the DAG in panel (a) of figure 2.4 as a useful definition of gender or sex discrimination in hiring, it does not provide an answer to the question of how to define discrimination more generally nor does it provide a solution to the problem of how to define discrimination in particular situations other than hiring discrimination—for instance, discrimination in track recommendations or grading at the end of elementary school in Germany.

In conclusion, a definition of discrimination as direct effect is only as general as the description of what constitutes the indirect effect(s). The most widely suggested general concept for such a definition is merit—similar, but certainly neither equivalent to qualifications or productivity nor to educational achievement without further arguments. Definitions of discrimination as unequal treatment conditional on merit I discuss below in section 2.4.4. They have several problems and limitations; the major problem is that they, too, cannot be developed without reference to contexts—e.g., labor market, education systems—and different variables that may or may not be part of merit in particular and of the model in general. For all the reasons given in this section, I will not adopt conceptualizations of discrimination as direct effect.

2.3.3 Perceptions, Beliefs, Information, and Signals

To circumvent the problems associated with causal effects of seemingly immutable characteristics such as race or sex when defining discrimination, some authors have turned to *perceptions* of (Fienberg & Haviland, 2003; Greiner & Rubin, 2010), *beliefs* (Angrist & Pischke, 2009, p. 5) and *information* (Berk, 2004, p. 96) about, or *signals* (M. Sen & Wasow, 2016) sent by the characteristics of interest.

Manipulability & well-defined causal questions

While I find D. B. Rubin (1986) to be the first to suggest that sex discrimination could be conceptualized as causal effect of an information or signal by "replacing an 'F' with an 'M' on a job application form" (D. B. Rubin, 1986, p. 962), Greiner and Rubin (2010, p. 776) credit Fienberg and Haviland (2003) with being the first to explicitly discuss discrimination as causal effect of perceptions. In a section of their comment on Pearl (2003) entitled "What is discrimination?", Fienberg and Haviland (2003) write:

> Discrimination is usually taken to mean the differential treatment of individuals based on a perceived characteristic or group membership. (Fienberg & Haviland, 2003, p. 319)

Fienberg and Haviland (2003) turn to perceptions and information for defining discrimination as a causal effect, since only at this level, so they argue, seemingly immutable characteristics, also called "concomitant variables" (Freedman, 2004, p. 283; Fienberg & Haviland, 2003, p. 319), such as sex or race are manipulable—be it through randomly allocated perceptions of different characteristics or by making information about a particular characteristic available versus not available.[9]

Berk (2004, pp. 82–84, 90–97), too, requires causes to be manipulable variables that can be changed by intervention and does not see attributes such as race or gender as such variables. His solution is essentially the same as in D. B. Rubin (1986), Fienberg and Haviland (2003), namely

> [...] to reformulate the intervention so that causal effects make sense. In the case of race, for instance, one can manipulate information about race, if not race itself. Thus, a job application could be doctored to show that the job applicant was white or black. (Berk, 2004, p. 96)

In a similar vein, Angrist and Pischke (2009, p. 5) define discrimination as causal effects of beliefs to enable precise causal questions that could potentially be answered by an experiment:

> [T]he issue economists care most about in the realm of race and sex, labor market discrimination, turns on whether someone treats you differently because they *believe* you to be black or white, male or female. (Angrist & Pischke, 2009, p. 5, their emphasis)

9 For the question who is to credit with with what exactly, it seems noteworthy that Fienberg and Haviland (2003) cite a working paper version of Bertrand and Mullainathan (2004) from 2003 as an example of such a strategy; they do not cite D. B. Rubin (1986).

In their introduction to *Counterfactuals and Causal Inference*, Morgan and Winship (2015) rely on perceptions in their definition of discrimination, to show that objections to the counterfactual approach are often misguided:

> [I]f discrimination is the topic of study, the attributes of individuals do not need to be manipulated, only the perception of them by potential discriminators. (Morgan & Winship, 2015, p. 440)

Also in Pearl et al. (2016) can be found an account of defining discrimination as the causal effect of perceptions. The authors discuss an example of discrimination in hiring on the basis of sexual orientation in which "Y stands for Mary's hiring, and X stands for the interviewer's perception of Mary's sexual orientation" (Pearl et al., 2016, p. 114). They then stress that

> [X] is the interviewer's *perception* of Mary's sexuality orientation, not the orientation itself, because an intervention on perception is quite simple in this case—we need only to imagine that Mary never mentioned that she is gay. (Pearl et al., 2016, p. 114, their emphasis)

In sum, *straightforward manipulability* and *well-defined causal questions* are the two arguments brought forward most often by authors from the causal effects literature to turn to perceptions or beliefs in defining causal effects of so-called immutable characteristics such sex or race in general and in defining discrimination in particular (Angrist & Pischke, 2009; Berk, 2004; Fienberg & Haviland, 2003).

However, both arguments have their downsides: First, it has been argued that the manipulability of perceptions might not be so straightforward after all. Indeed, perceptions and beliefs are within the mind of the perceiver and, thus, neither directly observable nor directly manipulable (Greiner & Rubin, 2010, p. 779; M. Sen & Wasow, 2016, p. 509). Relatedly, manipulations of perceptions do not always seem to be more realistic than manipulations of, for example, chromosomes. Therefore, M. Sen and Wasow (2016) suggest to turn away from perceptions to cues and signals—instead of, for instance, "perceived race", they suggest to investigate the causal effects of racial cues or signals.

Secondly, well-defined questions, or, synonymously, the fine articulation of causal states (Morgan & Winship, 2015, p. 38) does not hinge upon defining discrimination as the causal effect of perceptions, beliefs, or signals, for that matter. In fact, conceptualizations that rely on interventions at conception may be very precise—albeit unrealistic. So, straightforward manipulability and well-defined causal questions *alone* do not warrant a turn to either perceptions and beliefs or signals (M. Sen & Wasow, 2016).

Essentialist versus constructivistic views

A slightly less technical but more substantive argument in support of a turn from "immutable characteristics" such as sex or race to perceptions of such variables in defining discrimination, is that the "immutable characteristics" view of race or sex would imply a "biological definition" (Greiner & Rubin, 2010, p. 776) of or a "primordialist or essentialist" (M. Sen & Wasow, 2016) perspective on these characteristics. M. Sen and Wasow (2016), who focus on race, contrast this view with a constructivist framework "in which race is conceptualized as a complex, socially constructed identity with many mutable facets" (M. Sen & Wasow, 2016, p. 500). They argue that a constructivist perspective is superior to an essentialist perspective since

> Conceptualizing race and ethnicity in constructivist terms allows race to be disaggregated into constitutive elements, some of which can be manipulated experimentally or changed through other types of interventions. (M. Sen & Wasow, 2016, p. 500)

This way, the constructivist approach would, in contrast to the essentialist approach, allow to integrate seemingly immutable characteristics into the potential outcomes or counterfactual model of causality.

I agree with some important qualifications: While, of course, one has to allow that race is signaled and perceived in ways that allow prejudice, stereotypes, and other cognitive and affective mechanisms to work on the perception, it is by no means necessary to ignore or deny biological or genetic differences between human races (e.g, in skin pigmentation)—unless, of course, the definition of "race" refers to social and cultural dimensions only. Also, asking for the causal effect of an intervention at conception does not imply the assumption that cultural and social forces do not affect the outcome of interest, including the social construction and perception of a group of people. Note that both the discussion and my critique do apply for sex or gender, respectively. However, social class, a dimension of stratification and inequality I also care about in this study, is different and while it does not make any sense to conceptualize social class as a biologically fixed entity, it is, conversely, by no means necessary—and, actually, rather uncommon—to exclusively ask for the causal effect of perceptions of social class.

Total versus direct effects

That defining discrimination via perceptions may also solve the total effect problem discussed above is also an argument made in several contributions (Greiner & Rubin,

2010; M. Sen & Wasow, 2016; VanderWeele & Hernán, 2012). Unfortunately, most authors do not specify the kind of effect—total or direct—of perceived characteristics or signals when defining discrimination. Since it is the default causal effect, I suppose, most of them mean the total effect of an intervention that changes the signal or perception. Interestingly, VanderWeele and Hernán (2012), for example, explicitly define discrimination as direct effect of perceptions:

> Discrimination, on the other hand, is essentially the direct effect of sex controlling for all other variables at the time in which sex is perceived. (VanderWeele & Hernán, 2012, pp. 109–110)

However, a direct effect conceptualization of perceptions seems to be even more problematic than direct effect conceptualizations of attributes treated as immutable. In case of perceptions, a straightforward question to ask and examine would be whether the perceivers' prejudices or stereotypes mediate the effect of perceptions on the outcome. Now, if either or both variables would be included in such an analysis—to control for "all other variables" (VanderWeele & Hernán, 2012, p. 110)—the remaining direct effect of perceptions on whatever outcome would be net of the effect that is mediated by these variables—but would this direct effect really be an unbiased estimate of what we think of as discrimination? Probably not. In general, all arguments against direct effect conceptualizations from above apply and, therefore, I will certainly not define discrimination as the direct effect of perceptions, beliefs, or signals, respectively.

Conclusion *or* towards a useful definition of discrimination as a causal effect

While definitions of discrimination that rely on perceptions and beliefs are a considerable improvement over more traditional and imprecise definitions, they, too, have limitations. In fact, it could be argued that their major limitation resembles the manipulability problem the traditional definitions are accused of suffering from. If perceptions and beliefs are conceptualized as treatments, the relevant treatment lies within the mind of the perceiver and is, thus, neither directly observable nor directly manipulable (Greiner & Rubin, 2010, p. 779; M. Sen & Wasow, 2016, p. 509). This is not true for signals or, more generally, information. Another limitation of defining discrimination via perceptions or beliefs but also signals and information is that such definitions do not by themselves assure that they make more substantive sense than traditional definitions relying on so-called immutable characteristics. For example, direct effect conceptualizations of discrimination as causal effect of perceptions, be-

liefs, signals, or information suffer from the the same problem as other direct effect conceptualizations.

However, there are important advantages: While I think that the manipulability problem of so-called immutable characteristics could in many cases be solved by a little more imagination, the argument that causal effects of perceptions, signals, and the like, are easier to define is certainly true. This is particularly true for signals and information. Now, the major advantage of definitions of discrimination that rely on perceptions, beliefs, signals, or information is that they are substantively meaningful definitions even when conceptualized in terms of the preferred total effect. To me, this is the *key argument* for preferring definitions that feature perceptions, beliefs, signals, or information as treatments over traditional definitions relying on direct effects of so-called immutable characteristics or rather imprecise definitions such as those by Blank et al. (2004) and Quillian (2006) cited above. For reasons given above, I prefer definitions of discrimination that conceptualize signals or information as treatment over those that rely on perceptions and beliefs.

Based on the discussion in section 2.3, I propose the following general definition of discrimination: *Discrimination is the causal effect of an information about or a signal sent out by an individual on how this individual is treated by another individual.* Obviously, based on this definition, any human decision that distinguishes between individuals or groups of individuals constitutes discrimination. While this seems to be a limitation, it really is not. Certainly, the general definition can only be a starting point of any study of discrimination; what the researcher has to do, is to specify which information or signal is of interest and, thus, which special type of discrimination is under study. In the present study, I am interested in discrimination by teachers based on information about students' ethnicity, social class, and sex—also known as *ethnic* discrimination, *social class* discrimination and *sex* or *gender* discrimination. Defining these special forms of discrimination requires to adapt the general definition accordingly. In the case of ethnic discrimination, the definition would read: *Ethnic discrimination is the causal effect of an ethnic information about or an ethnic signal sent out by an individual on how this individual is treated by another individual.* As soon as there is a causal effect of an ethnic signal on the treatment of an individual by another individual, for example the treatment of a student by a teacher, there is ethnic discrimination.

Above I frequently referred to counterfactual questions as a tool to understand causal effects. My general definition of discrimination requires the following counterfactual question to be answered: How would an individual, *i*, have treated another individual, *j*, had an information about or a signal *s* sent by individual *j* been different? To establish discrimination, this question needs to be answered with "differently",

which only highlights how general the definition is. To establish a particular type of discrimination, it has to be specified which information or signal *s* is of interest and which are the alternative causal states under investigation. In case of ethnic discrimination, the general question might be rephrased as follows: How would an individual, *i*, have treated another individual, *j*, had signal *s* sent by individual *j* signaled a Turkish background instead of German background? Note further that under this definition any special form of discrimination cannot occur in the absence of the corresponding information or signal and such a form of discrimination cannot be established without variation in *s*. Ethnic discrimination, for example, cannot occur in the absence of an ethnic information or signal and it cannot be established without variation in the ethnic nature of the signals sent by individuals.

I would like to conclude with an example from discrimination in education and a reminder: Decisions by teachers that disadvantage a single student or a group of students of particular ethnic background in the aggregate are not necessarily an example of ethnic discrimination. Instead, this may very well constitute social class discrimination, for example, or, really, discrimination based on merit, achievement, or performance. *That* these different forms of discrimination *are judged* rather differently in contemporary societies should be obvious. *How to judge* these and other forms of discrimination cannot be answered using methods empirical social scientists have at their disposal.

2.4 Other Conceptualizations of Discrimination in the Social Sciences and Beyond

2.4.1 Discrimination at the Individual Versus Group Level

In chapter 1 I have argued that there are two major reasons for studying discrimination: First, many forms of discrimination—for instance, by virtue of a person's race or ethnicity, gender, or social class—"violate our society's sense of fairness" (Holzer & Ludwig, 2003, p. 1148). Secondly, discrimination is often motivated as a potential explanation for disparities and inequalities between different societal groups. I argue that these motivations relate to a distinction highlighted mainly by economists, namely the distinction between *discrimination at the individual level* and *discrimination at the group level* (Heckman, 1998, p. 102).

Becker (1957/1971): Individual versus market discrimination

Heckman (1998) credits Becker (1957/1971) with being the first to observe that individual discrimination does not necessarily aggregate to what has been called market discrimination (Becker, 1957/1971, pp. 17–18, 43–45, 84–100):

> It was Becker's (1957) insight to observe finding a discriminatory effect of race or gender at a randomly selected firm does not provide an accurate measure of the discrimination that takes place in the market as a whole. (Heckman, 1998, p. 102)

Whether and to what extent individual preferences of, for instance, employers turn into discriminatory behavior and aggregate to market discrimination against a particular group of minority workers, in turn, depends on multiple market forces and conditions. According to Becker (1957/1971), the most important factors are the size of the minority group and the distribution and the valence of employers' preferences—not the preferences "of an average or 'representative' employer" (Becker, 1957/1971, p. 43). Becker (1957/1971)'s arguments are nicely summarized by Heckman (1998):

> The impact of market discrimination is not determined by the most discriminatory participants in the market, or even by the average level of discrimination among firms, but rather by the level of discrimination at the firms where ethnic minorities or women actually end up buying, working and borrowing. (Heckman, 1998, p. 102)

While this might seem to be a trivial insight, it is fundamental for designing, conducting, and analyzing data from audit and correspondence studies—but usually ignored (Heckman, 1998).

Note that Becker (1957/1971)'s discussion is not concerned with the conditions of how individual preferences turn into behavior. The insight that prejudice and discrimination are two different things and that prejudiced actors do not necessarily discriminate according to their prejudices is much older (LaPiere, 1934; Merton, 1949) (also see section 2.2.1). What Becker (1957/1971) adds is a systematic and formal discussion of discrimination in the market as a whole. For this, he certainly deserves credit. I will talk about Becker (1957/1971)'s theory of taste discrimination and how it applies to education in more detail below in chapter 3. How the insights by Becker (1957/1971) affect the design of experimental studies on discrimination I discuss in chapter 6.

Individual and group discrimination in models of statistical discrimination

In contrast to Becker (1957/1971), who argues that not every actor who would discriminate in principle also gets the opportunity to do so, Aigner and Cain (1977) show that in some models of *statistical discrimination* individual discrimination that actually takes place does not aggregate to group discrimination, which

> …in labor markets is evident when the average wage of a group is not proportional to its average productivity. (Aigner & Cain, 1977, p. 178)

Now, following statistical discrimination theory, employers lack perfect knowledge about the true productivity, q, of a candidate and, therefore, construct a weighted average of observed individual performance, y, and group productivity, α, to come up with an estimate, $\hat{q}$:

$$\hat{q} = (1 - \gamma)\alpha + \gamma y \tag{2.4}$$

If the observed performance, y, is indeed an imperfect measure of the true productivity, q—i.e., if $\gamma < 1$—and α differs between two groups, A and B, so that, e.g., $\alpha^A > \alpha^B$, candidates who are (or, as I would like to add based on the discussion in section 2.3.3, appear to be) members of group B will estimated to be of lower productivity than had they been (or appeared to be) members of group A. Clearly, this is an account of discrimination as an individual-level causal effect by virtue of membership in group A versus B.

However, note that both groups, A and B, will—on average—be paid according to average productivity as long as the decision maker gets α^A and α^B right and some additional assumptions hold (Aigner & Cain, 1977), since $E[(1-\gamma)\alpha+\gamma y] = \alpha$. That is, as long as decision makers rely on correct bliefs about average group productivity, there is no group discrimination in the sense of Aigner and Cain (1977). So, even if every individual member of group B suffers from individual discrimination in the sense of a causal effect as described above, the group as a whole will not be disadvantaged.

While to some, this result might still be counterintuitive and the assumption that employers hold correct beliefs about the average productivity of different groups might seem strong, the discussion in Aigner and Cain (1977) shows that it is not necessary to turn to market forces to find that individual discrimination is not equivalent with group discrimination—that is, a disadvantage on the group level. Therefore, Aigner and Cain (1977)'s discussion certainly adds to Becker (1957/1971)'s insight that explaining inequality through discrimination is everything but straightforward. I provide a more elaborate discussion of different models of statistical discrimination in section 3.1.2.

Group discrimination or inequality?

I find the distinction between individual discrimination and group discrimination to be crucial for all researchers who study inequality between groups. To explain inequality through discrimination, it is necessary that individual discrimination aggregates to group discrimination. However, group discrimination should not be confused with inequality. These phenomena, albeit related, are not equivalent and the existence or direction of one of them does not say anything about the existence or direction of the other. Obviously, it is possible that there is no group discrimination but inequality between groups. This is just a different way of stating the obvious: There are other causes of inequality between groups than discrimination.

Maybe less obvious on first sight is that group discrimination does not necessarily lead to inequality between groups. An example for such a scenario would be a successful affirmative action policy, where group discrimination against group A offsets inequality that initially existed to the disadvantage of group B but wasn't due to discrimination against group B. Following the same logic, it is certainly also possible that there is group discrimination against group A but inequality to the disadvantage of group B. This happens, for instance, when affirmative action policies that discriminate against group A are not successfully equalizing the disadvantage of group B that, again, was not due to discrimination against group B.

2.4.2 Definitions of Discrimination Based on Group Membership

Defining discrimination as differential treatment of individuals based on their group membership instead of their individual characteristics is quite common. Take this definition by Levin and Levin (1982, p. 51):

> Discrimination can be defined as differential or unequal treatment of the members of some group or category on the basis of their group membership rather than on the basis of their individual qualities.

In the same vein, Allport (1954, pp. 51–52) explicitly rules out individual qualities as causes of discrimination: "Differential treatment based on individual qualities probably should not be classed as discrimination". More examples of definitions of discrimination in reference to group membership can be found in Allport (1954), Dovidio et al. (2010), Fishbein (2002), Nelson (2006), and others.

While, again, conceptualizations of discrimination often refer to group membership, I suggest that such an approach is actually not very helpful. The key argument here is that "[a]ll individual characteristics define groups" (England & Lewin, 1989,

p. 241). This also applies to characteristics that determine or are correlated with what is called, among others, merit, productivity, or achievement—e.g., test scores (Aigner & Cain, 1977; England & Lewin, 1989). Proponents of definitions of discrimination based on group membership typically see such characteristics as individual. However, and as England and Lewin (1989) argue, individuals with the same or similar "individual" characteristics could always be grouped according to this characteristic. An individual student that scores beyond average on a particular achievement test belongs to the group of students that score beyond average on that test. An individual student that has completed a particular extracurricular activity in school belongs to the group of students that completed this extracurricular activity. And so on. But not only do all individual characteristics define groups, also can all group characteristics be reduced to individual characteristics—just take the examples of skin color and religion, which are, under such definitions of discrimination, typically considered to be group characteristics: As a matter of fact, the color of an individual's skin is this individual's skin color as the religious beliefs of an individual are this individual's religious beliefs. And so on.

However, I am not arguing that the act of discriminating between two individuals based on whatever characteristic is unrelated to the distinction of individual and group. In fact, I suggest that both the act of discriminating and research on discrimination alike rely on this distinction and, in particular, on the group concept or, more broadly, the concepts of category and categorization. I discuss different mechanisms of why and how people discriminate in general and on particular grounds in particular in more detail in chapter 3. Here, I only want to briefly point to the importance of the act of categorizing or grouping people for how they are treated: Stereotypes and prejudice, key determinants of discriminatory behavior and connected to groups of people by definition, are themselves—i.e., their contents as well as their application—influenced by the perception of whether or not we see ourselves and others as part of particular groups or not (see, e.g., Bless et al., 2004; Bless & Schwarz, 2010; Fiske, 1993b; Fiske et al., 1999; Macrae & Bodenhausen, 2000; Tajfel, 1970; Tajfel & Turner, 1986, among many other contributions from research in social cognition).

Put differently, how individuals are categorized based on the characteristics they possess determines how they are treated. This is, because, according to the research in social cognition cited above, the human brain links these categories to knowledge and expectations, that is, stereotypes, schemata, scripts—as well as to affects and emotions, that is, prejudice. And while virtually all theories of discrimination rely on the notion of group membership, the definition of the phenomenon that they seek to explain does not have to—and, as I argue, must not—rely on this concept.

On the omnipresence of group information

One might even go so far as to doubt that any act of treating a person can be purely individual. While the input and, thus, the values on the variables used by the actor might only stem from the target—i.e., no group averages or other aggregated information is used as a variable—the question of how to weight the different input factors seems virtually impossible to address without referring to group information or information that is aggregated over individuals, respectively.

Imagine, a teacher wants to track students based on a calculation of success probabilities on different secondary school tracks. Imagine further that the teacher has all the information at hand that, theoretically, would suffice to (almost) perfectly predict future success on different tracks. Now, the question is how the teacher weights the different factors or variables to calculate the probability. To me it seems impossible to do so without drawing on either personal experience with other students or other teachers' experiences with other students or other more or less correct knowledge about what students that are similar or different to the student in question are capable of. Technically speaking, any model—be it a statistical model set up by a researcher or an unconscious mental algorithm an actor's brain relies on—that is meant to predict any outcome has to rely on data of more than just one unit to infer how data from individual level variables should be weighted. In fact, formal models of discrimination, such as statistical discrimination models (e.g., Aigner & Cain, 1977) are a nice case in point: Decision makers only know when to turn down group means and, thus, selecting candidates on "individual" information only, when they have information on the individual information's reliability that, in turn, can only be calculated from group level or aggregate level data. In sum, the problem of imagining a judgment that does not contain *any* knowledge about groups is yet another reason, why—in a definition of discrimination—the distinction of individual and group characteristics does not seem very useful to me.

But even if one does not share this fundamental skepticism laid out in the last paragraphs, I think that the point against definitions of discrimination based on the distinction between individual and group characteristics has been convincingly made by England and Lewin (1989). Of course, sociologists that study discrimination are and will be interested in discrimination by virtue of characteristics that form groups that are of societal or political relevance such as immigrants, racial or ethnic groups, social classes, or men and women. This is because the analytical focus of sociology is on the collective level or macro level, respectively. This, however, is no good reason to build a definition of discrimination around the group concept. In sum, based on England and Lewin (1989), I argue that a definition of discrimination that relies on the dis-

tinction between group characteristics and individual characteristics is meaningless and, thus, I will neither adopt such definitions, nor will I conceptualize discrimination based on this distinction.

2.4.3 Definitions of Discrimination Based on the Distinction Between Ascription and Achievement

Some authors turn to the concepts of ascription and achievement to distinguish between non-discriminatory behavior and discriminatory behavior. England and Lewin (1989), for instance, call a treatment discriminatory if and only if it is based on "ascriptive group memberships" (p. 239) or "ascriptive statuses" (p. 241). Studies on discrimination in German education have also built their definitions around the concept of ascription (Diehl & Fick, 2016; Kalter, 2003; Kristen, 2006a).

The history of the terms ascription and achievement dates back to Linton (1936) who coined them in his anthropological "Study of Man". He distinguished between "two types of statuses, the ascribed and the achieved" (Linton, 1936, p. 115):

> *Ascribed* statuses are those which are assigned to individuals without reference to their innate differences or abilities. They can be predicted and trained for from the moment of birth. The *achieved* statuses are, as a minimum, those requiring special qualities, although they are not necessarily limited to these. They are not assigned to individuals from birth but are left open to be filled through competition and individual effort. (Linton, 1936, p. 115, his emphasis)

The distinction was quickly picked up by sociologists—it became central to Parsons (1940, 1950)'s scheme and analysis of stratification, for example. Young (1958), to whom I will return in the next section, saw the turn from rewarding ascription to rewarding achievement as one major symptom and mechanism of the "Rise of the Meritocracy" in the industrialized and modern world.

Even though many sociologists have adopted it, I suggest that the distinction between ascription and achievement is not very fruitful in research on discrimination and that definitions of discrimination should not be based on it. I think that the distinction does not clarify but rather confuse the situation: What does the distinction add to the general definition of discrimination I proposed above in section 2.3.3? Typically, researchers would—using either definition—specify on the basis of which characteristic or signal discrimination is meant to occur. Using the general definition I advocate in the present study, the researcher can just carry on, once this has been stated. However, the researcher that conceptualizes discrimination as differential treatment

by virtue of ascribed characteristics only, has to make sure that the characteristic in question is truly ascribed. But which characteristic is truly ascribed, I would argue, is not always clear. Remarkably, most authors neither explain the distinction between ascribed and achieved characteristics nor do they cite a source that does—typically, examples of ascribed characteristics are provided (e.g., England & Lewin, 1989; Reskin, 2003).

Take the example of attractiveness: In the labor market there are certainly jobs for which there are positive returns for attractiveness (e.g., Hamermesh & Biddle, 1993; Jæger, 2011; López Bóo et al., 2013; Wong & Penner, 2016): A more attractive person might be preferred over a less attractive person in a hiring process since the employer knows that on the position to be filled a beautiful person will, ceteris paribus, be more productive—e.g., because customers prefer to buy from more attractive sellers. Now, differences in attractiveness are certainly partly "innate differences" and can partly be achieved though individual effort—both are named by Linton (1936) as mechanisms of how achieved statuses are assigned. Also, the hiring process is competitive. Does that mean that attractiveness is an achieved characteristic? If yes, should this really be the reason for not calling this hiring process to constitute an example of discrimination by virtue of attractiveness in a particular labor market? What if attractiveness is not related to productivity but used by employers to distinguish among candidates nevertheless? Does that make it an ascribed characteristic? Or is it an ascribed characteristic because it is "assigned to individuals from birth" (Linton, 1936), which is the definition of ascribed status that might simply be used to define ascribed characteristics?

It seems clear to me that the terms ascription and achievement are typically used to suggest that some kind of inequality—for example, unequal treatment—is unfair or unjust because it is based on so called ascriptive characteristics. In contrast, differential treatment that is based on so called achievement is not called discrimination since it is considered fair, just, or at least not bad. However, quite obviously, such judgments are normative and, hence, should be avoided in an empirical study. In conclusion, I think that the distinction between ascription and achievement adds nothing to the definition or empirical study of discrimination at all. Why we study discrimination on the basis of some characteristics more often than on the basis of others might certainly be linked to value judgments in the society and their understanding of what constitutes fair and unfair processes or outcomes. However, to build a definition of discrimination around this distinction makes no sense to me.

2.4.4 Definitions Based on Merit

Defining discrimination as unequal or differential treatment conditional on *merit* has a long history in anti-discrimination law (DeSario, 2003; McCrudden, 1998) and has also found its way into the empirical social science literature (e.g., Driessen et al., 2008). However, conceptualizing discrimination in such a way bears the problem of having to define merit. But how should merit be defined? In principle, this question is no different to the question of how to define discrimination. There is no true definition of merit, although, of course, there are historical and famous examples of what is meant by merit.

Young (1958), who is typically credited with coining the term *meritocracy*, defined merit as intelligence plus effort: "Intelligence and effort together make up merit ($I + E = M$)" (Young, 1958). But is this a complete definition? And how should merit be distinguished from, say, productivity in the labor market or the likelihood of success in the education system? Is the parental support that students from higher social classes receive to a greater extent and that makes them more likely to pass exams with good grades part of their merit? According to Young (1958, p.94), probably not. Hence, for identifying discrimination defined as inequality conditional on merit, we would have to condition on parental support, wouldn't we? Based on the approach I have taken and discussed extensively in section 2.3 that focuses on the causal effect of behavior: Certainly not. However, other approaches might give affirmative answers. One is the institutional discrimination approach that I discuss in greater detail in section 2.4.7 below.

So, I second A. Sen (1999) in his description of the problems of the terms merit and meritocracy:

> The idea of meritocracy may have many virtues, but clarity is not one of them. The lack of clarity may relate to the fact [...] that the concept of "merit" is deeply contingent on our views of a good society. (A. Sen, 1999, p. 5)

This is a major problem that makes the term merit and, thus, a definition of discrimination that is based on it, susceptible to redefining it for ideological and political reasons—which is precisely what people, including scholars, tend to do (Quillian, 2006; Uhlmann & Cohen, 2005). In conclusion, I object to definitions of discrimination based on merit for the same reasons I object to definitions of discrimination as unfair treatment: They require normative judgments and facilitate normative interpretations of findings on discrimination. Such definitions I will not adopt.

2.4.5 Differential Treatment Versus Differential Impact

In their widely cited report, Blank et al. (2004) define discrimination as follows:

> (1) *differential treatment on the basis of race* that disadvantages a racial group and (2) *treatment on the basis of inadequately justified factors other than race* that disadvantages a racial group (differential effect). (Blank et al., 2004, 39, their italics)

Blank et al. (2004, pp. 39–42) further explain that both components of their definition are "based on behavior or treatment that disadvantages one racial group over another" (p. 39). However, racial discrimination in the sense of the first component—i.e., *differential treatment*—"occurs when a member of one racial group is treated less favorable than a similarly situated member of another racial group and suffers adverse or negative consequences" (p. 40). "Intentional discrimination of this kind", Blank et al. (2004, p. 40) add, would typically be unlawful in the US in many areas such as employment, housing, and education.

The first component of this definition has similarities with definitions discussed above, namely the conceptualization of discrimination as ceteris paribus causal effect (Heckman, 1998), and also with the definition of racial discrimination as causal effect of race in the counterfactual sense—which is also discussed by Blank et al. (2004, chapter 5). However, I have issues with this first component and the way Blank et al. (2004) describe it. First, the authors confuse individual and group level, when they depict discrimination as "differential treatment on the basis of race that disadvantages a racial group", suggesting that a treatment does not count as discrimination when it does not disadvantage a group—as a whole or on average. As discussed in 2.4.1, however, discrimination on the individual level does not readily aggregate into discrimination on the group level. In their comment afterwards they refer to "a member of one racial group" (p. 40) that "suffers adverse or negative consequences" (p. 40) but no longer to the whole group. Also, from their discussion of statistical discrimination (pp. 61–63) and discrimination as a causal effect (pp. 77–81), I conclude that they—in contrast to their definition—do not really restrict discrimination to treatment that necessarily results in a disadvantage on the group level.

Secondly, Blank et al. (2004) refer to the first component of their definition as "[i]ntentional discrimination" (p.40). Obviously, such a restriction would rule out many forms of discrimination that are unintentional in the sense that they are rooted in unconscious or implicit cognitive or motivational processes the actor is not aware of. However, when they discuss similarities and differences of their definition with the conceptualization of disparate treatment in US law and jurisprudence, they say that

subtle forms of discrimination that are "perhaps unintentional" would "fall within the scope" of their definition (Blank et al., 2004, p. 41).

The second component—called *differential effect discrimination* by Blank et al. (2004, p. 39)—is related to what is known as *disparate impact discrimination* in US law. I find this second component even more problematic than the first component, mainly for a reason discussed above in section 2.2.3 and other sections: there is no scientific method to determine what inadequately justified factors are—or, for that matter, what adequately justified factors are. Also, from the discussion in Blank et al. (2004) it is not clear to me, what *they* suggest these factors are.

With regard to the second component, my conclusion from a social science perspective is that there are two alternatives of how to proceed: First, we dismiss the condition of "inadequately justified factors" and, thus, conceptualize differential effect discrimination as treatment on the basis of any factor other than the factor in question—e.g., race in a study on racial discrimination; gender in a study on gender discrimination—that disadvantages a member of the group in question (individual discrimination) or the group as a whole (group discrimination). Note that such a conceptualization would inevitably lead to a situation in which every employment or tracking decision would be an instance of discrimination. Whatever mechanism is implemented, discrimination would be the result as long as members of a particular group are preferred over members of another group—for instance, because they perform better on a standardized test. Note that I have argued that this is indeed the most general conceptualization of discrimination—but not of racial discrimination—and, in fact, a very useful point to start a discussion on how to conceptualize discrimination. However, it is not a good point to end such a discussion. It is not, because, defined in this way, discrimination would not be different from inequality between groups.

Secondly, we could forget about differential effect discrimination altogether and examine such processes under the label inequality. Eventually, from a social science perspective, it really does not matter, whether we study such processes under the label discrimination or under the label inequality, since definitions "do nothing but introduce new arbitrary shorthand labels; they cut a long story short" (Popper, 1945). However, it is in the long stories where our research interests lie.

2.4.6 Disparate Treatment Versus Disparate Impact

> To study the social scientific definition of discrimination, one must attend to the legal definition of discrimination.
>
> *(Lucas, 2008)*

The two components of the definition proposed by Blank et al. (2004)—differential treatment and differential effect discrimination—discussed in section 2.4.5 are based on the concepts of *disparate treatment discrimination* and *disparate impact discrimination* as developed in US law and jurisprudence.

Section 703 of title VII of the Civil Rights Act from 1964 declares it "an unlawful employment practice" …

> (1) to fail or refuse to hire or to discharge any individual, or otherwise to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such *individual's race, color, religion, sex, or national origin*; or
>
> (2) to limit, segregate, or classify his employees in any way which would deprive or tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee, because of such *individual's race, color, religion, sex, or national origin*.

Disparate treatment discrimination in US law

US jurisprudence and legal scholars from the US have interpreted this section as declaring unlawful two different forms of discrimination, namely *disparate treatment discrimination* and *disparate impact discrimination*. The former, disparate treatment discrimination, has typically been interpreted by US courts as intentional discrimination that the defendant engages in consciously. Green (2003) summarizes what he calls "traditional disparate treatment theory" as follows:

> Disparate treatment doctrine has long been understood to require a showing of *intentional discrimination*, often defined in terms of *conscious motivation to discriminate*. (Green, 2003, 113, my italics)

In a similar vein, Krieger and Fiske (2006) summarize the underlying assumptions of antidiscrimination laws and the related jurisprudence as follows:

> In numerous ways, antidiscrimination law reflects and reifies a commonsense theory of social perception and judgment that attributes disparate treatment discrimination to the deliberate, conscious, and intentional actions of invidiously motivated actors. (Krieger & Fiske, 2006, p. 1028)

These unrealistic assumptions about human cognition in general and about the mechanisms of how stereotypes, prejudice, and discrimination work in particular have given rise to critique from social scientists and legal scholars alike (Greenwald & Krieger, 2006; Krieger, 1995; Krieger & Fiske, 2006; Oppenheimer, 1993). I second this critique, because neither social science nor legal definitions of discrimination should be based on false premises (also see my discussion in section 2.4.5).

Disparate impact discrimination in US law

Disparate impact discrimination has been even more controversial than disparate treatment discrimination. Sometimes, sentence 2 cited above has been "identified as the source of the theory of disparate impact" (Rutherglen, 1987, p. 1300).

Disparate impact discrimination theory recognizes indirect or subconscious forms of discrimination that are not directly linked to an employee's individual's race, color, religion, sex, or national origin but have nevertheless adverse consequences for employees with such characteristics. In the 1971 case of Griggs v. Duke Power Co., the Supreme Court of the United States established the disparate impact doctrine that under which it is prohibited to apply "facially neutral employment practices with a disproportionately adverse effect on protected groups, even in the absence of discriminatory intent" (DeSario, 2003, p. 480), "unless [the employer] can show a business justification for those practices" (Peresie, 2009, p. 776).

Rutherglen (1987, p. 1297) sees disparate impact discrimination as "the single most important judicial contribution to title VII of the Civil Rights Act of 1964" and praises it—compared to a theory of intentional discrimination, that is disparate treatment discrimination—as "an objective theory of discrimination" (Rutherglen, 1987, p. 1298). Concerning the differential impact definition of discrimination by Blank et al. (2004), I have argued exactly the opposite in section 2.4.5 above and will stick to my interpretation. Rutherglen (1987) argues differently, since he compares disparate impact discrimination to disparate treatment discrimination in the sense of intentional and conscious discrimination that "requires a finding about the defendant's state of mind" (p.1298). However, I have already rejected this reading of disparate treatment discrimination as ignoring the reality of discriminatory treatment and I also cannot see how practices that have adverse effects can objectively be justified through methods

available to empirical social scientists. Even legal scholars struggle over the question which practices are justified and which are not (see, e.g., DeSario, 2003, for a discussion on different conceptualizations of meritocracy that are used to justify such practices). Thus, from a social science perspective, I cannot follow Rutherglen (1987) in his praise.

For a more thorough discussion of legal definitions and conceptualizations of different forms of discrimination based on US law from a social science perspective, see, e.g., Blank et al. (2004), Greenwald and Krieger (2006), Krieger (1995), Krieger and Fiske (2006), Lucas (2008).

Discrimination in European and German law

A similar distinction has been introduced to European law by the Amsterdam Treaty and directive 2000/43 ([2000] OJ L180/22), sometimes called *Racial Equality Directive* (Bell, 2008, p. 36) or *Race Directive* (Zschirnt & Ruedin, 2016), and directive 2000/78 ([2000] OJ L303/16), sometimes called *Employment Equality Directive* (Bell, 2008, p. 36). The two directives call for a prohibition of discrimination on grounds of racial or ethnic origin as well as religion or belief, disability, age, and sexual orientation in employment and vocational training Bell (see, e.g., Bell, 2008).

The directives have been implemented into German law: Section 1 of the German General Act on Equal Treatment (dt. Allgemeines Gleichbehandlungsgesetz, AGG, 2006, §1, my italics), says:

> The purpose of this Act is to prevent or to stop discrimination on the grounds of race or ethnic origin, gender, religion or belief, disability, age or sexual orientation.

It then distinguishes between *direct discrimination* and *indirect discrimination* (AGG, 2006, §3, my italics):

> (1) *Direct discrimination* shall be taken to occur where one person is treated less favourably than another is, has been or would be treated in a comparable situation on any of the grounds referred to under Section 1.
>
> (2) *Indirect discrimination* shall be taken to occur where an apparently neutral provision, criterion or practice would put persons at a particular disadvantage compared with other persons on any of the grounds referred to under Section 1, unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary.

In general, the arguments I brought forward against US law as a basis of a social science definition of discrimination also apply to European and German law. Note that, interestingly, the word "discrimination" (dt. *Diskriminierung*) does not appear once in the original German text of the AGG (2006). With or without explicitly mentioning the word discrimination, definitions of discrimination based on antidiscrimination laws as currently implemented in the US, the EU, and Germany are not very useful for social science research.

2.4.7 Institutional, Structural, and Systemic Discrimination

The conceptualizations of discrimination I discuss in this section, characterize forms of discrimination according to the societal level on which *causes or mechanisms* of discrimination are theorized to lie. This is different to the distinction of individual and group discrimination from section 2.4.1 that relates to the level on which the *effects* of discrimination are measured.

Also, *institutional* discrimination, *structural* discrimination, and *systemic* discrimination are typically introduced as theoretical approaches that are supposed to help explain discriminatory treatment and persisting inequality between societal groups. However, the present chapter is not about theories and their mechanisms. Here, I will just briefly discuss the different conceptualizations of these forms of discrimination. As theories, I will discuss institutional, structural, and systemic discrimination in chapter 3. In this discussion I will also address the question, whether these theories can really be called proper theories.

In my discussion in the following section, I focus on the institutional form of discrimination (e.g., Sidanius & Pratto, 1999) and the original, more widely used, but also—in this more general study on discrimination in German education—less applicable term, namely institutional racism (e.g., Carmichael & Hamilton, 1967; J. M. Jones, 1972; L. L. Knowles & Prewitt, 1969) that I start my discussion with.

Note that some authors discuss institutional*ized* discrimination (e.g., Berard, 2008; J. R. Feagin, 1977; J. R. Feagin & Booher Feagin, 1986) and institutional*ized* racism (C. P. Jones, 2000). The terms *institutional bias* (e.g., Henry, 2010) and, much less often, *institutionalized bias* (e.g., Sundstrom, 1990) are also in use. Also note that, sometimes, the terms institutional, structural, and systemic are used interchangeably (e.g., J. R. Feagin & Bennefield, 2014).

Institutional racism

The activist Stokely Carmichael and the scholar Charles V Hamilton Carmichael and Hamilton (1967) are typically credited with having coined the term "institutional racism"—a term much more widely used than "institutional discrimination"—to contrast it with "individual racism" that was suggested to "consist of overt acts by individuals, which cause death, injury or the violent destruction of property" (Carmichael & Hamilton, 1967, p. 4). Institutional racism, in contrast, was introduced as

> […] less overt, far more subtle, less identifiable in terms of specific individuals committing the acts. But it is no less destructive of human life. [It] originates in the operation of established and respected forces in the society, and thus receives far less public condemnation […]. (Carmichael & Hamilton, 1967, p. 4)

While this description is both too vague and too empirical to be a useful definition for social science research, it conceptualizes institutional racism rather clearly as originating in the operation of established and respected forces in the society—that is, in a society's institutions.

The concept of institutional racism was then picked up, developed, and investigated further by L. L. Knowles and Prewitt (1969), who unfortunately fail to provide a clear definition. In contrast to Carmichael and Hamilton (1967), L. L. Knowles and Prewitt (1969, p. 15) explicitly discuss both unintentional and intentional forms of institutional racism. The widely cited definition by J. M. Jones (1972) is more clear, but introduces a normative connotation by referring to inequities instead, for instance, to the more neutral terms inequalities or disparities:

> Institutional racism can be defined as those established laws, customs, and practices which systematically reflect and produce racial inequities in American society […] *whether or not the individuals maintaining those practices have racist intentions.* (J. M. Jones, 1972, p. 131, his emphasis)

Similar to L. L. Knowles and Prewitt (1969), he allows institutional racism to be "either overt or covert […] and either intentional or unintentional" (J. M. Jones, 1972, p. 131).

Institutional discrimination

That the term "institutional racism" became more popular than the term "institutional discrimination" in the US literature and, thus, in the English speaking literature as a whole, is not surprising, given the history of and the, therefore fully comprehensible,

focus on *race* relations in the US American society. However, the term "institutional discrimination" seems to be older (e.g., Myrdal, 1944, pp. 606, 629, 631; Greene, 1938, p. 211)[10]. That discrimination may be institutionalized and, thus, may become itself an institution—not an organization—but also embedded in other institutions, e.g., the law, was also recognized by Antonovsky (1960).

However, accounts of explicit definitions of institutional or institutionalized discrimination are younger. As one would expect, definitions of institutional discrimination are typically broader in the sense that they apply not only to race but other dimensions such as social class or gender and more narrow in the sense that they focus on treatment and behavior and leave beliefs and attitudes aside. Studies on German education therefore usually use the term *institutionelle Diskriminierung* instead of *institutioneller Rassismus*[11].

An in-depth discussion and explicit definitions of two forms of institutional discrimination is provided by J. R. Feagin and Booher Feagin (1986):

> [...] *direct institutionalized discrimination* refers to organizationally-prescribed or community-prescribed actions which have an intentionally differential and negative impact on members of subordinate groups. (J. R. Feagin & Booher Feagin, 1986, p. 30)
>
> [...] *indirect institutionalized discrimination* refers to practices having a negative and differential impact on minorities and women even though the organizationally prescribed or community-prescribed norms or regulations guiding those actions were established, and are carried out, with no prejudice or no intent to harm lying immediately behind them. (J. R. Feagin & Booher Feagin, 1986, p. 31)

Obviously, the main difference between the two forms is that one refers to intentional and the other to unintentional behavior. Less obvious is what is meant by differential and negative impact in both definitions. However, from the discussion in J. R. Feagin and Booher Feagin (1986) but also in J. R. Feagin (1977), J. Feagin and Eckberg (1980) it is clear that the authors are interested not only—and, in fact, not even foremost—in discrimination in the sense of a causal effect of a signal of some sort, as defined in section 2.3.3, but in inequality more generally. Interestingly, according to both definitions, only "subordinate groups" and "minorities and women" may suffer from discrimination.

---

10 Note that, in some parts in the book, Myrdal (1944) uses the terms discrimination and segregation synonymously and discusses "institutional segregation" throughout the book.

11 *Institutionelle Diskriminierung* and *institutioneller Rassismus* are the literal German translations of institutional discrimination and institutional racism, respectively.

As for a final example let's have a look at a rather simple, more general, and, thus more useful definition of institutional discrimination. In their study on discrimination in US education, Meier et al. (1989) state:

> Institutional discrimination occurs when the norms, procedures, and rules of an organization discriminate against certain individuals. (Meier et al., 1989, p. 30)

Here, institutional discrimination is neither explicitly nor obviously equated with inequality. However, note that as long as it is not clear what is meant by discrimination, we cannot judge how useful such a definition really is. From the discussion in Meier et al. (1989) it seems that, indeed, discrimination in education is equated with inequality of educational opportunity based on race and social class. Note further that in this definition, institutional discrimination is a phenomenon occurring within organizations.

Contributions to the German literature

The contributions on the topic published in German tend not to provide definitions that are more useful—if they provide own self-contained definitions at all: Gomolla and Radtke (2010), the most widely cited study on institutional discrimination in German education, for instance, doesn't. In fact, the definitions provided in the German literature reflect the problems of their international counterparts: First, while it is not always specified what exactly is meant by *institutionell* or *Institution*, some define *institutionelle Diskriminierung* more generally as institutionalized or embedded in institutions in a general sense (Ditton & Aulinger, 2011, p. 102), some restrict it to processes within organizations (Gomolla, 2016, p. 2; Hasse & Schmidt, 2012, p. 886).

Secondly, all of them share an approach that defines discrimination through outcomes on the group level and tend to conflate discrimination with inequality—either conditional inequality, such as inequality of opportunities, but, quite frequently, also unconditional inequality, that is, inequality of outcomes. So, all too often, institutional discrimination is conceptualized as an effect on between-group inequality (Ditton & Aulinger, 2011; Gomolla, 2016), not on discrimination as phenomenon of interindividual behavior in the sense I advocate in this study (see section 2.3.3). Remarkably, even those proponents of institutional discrimination in the German debate that recognize this approach as problematic, do *not* adapt their definition accordingly (e.g., Gomolla, 2016, p. 12).

How useful are definitions of institutional discrimination?

In this section I will briefly summarize the issues I have with the definitions of institutional racism and discrimination. Remember that these issues are of conceptual and methodological nature, not of theoretical nature, and concern the usefulness for empirical research.

My first issue with many, not all, definitions of institutional racism or institutional discrimination is that they lack clarity regarding key terms. With regard to institutions, all too often it is unclear whether the term refers more generally to the "the rules of the game in a society" (North, 1990, p. 2) as in J. R. Feagin (1977), J. M. Jones (1972), or to tangible organizations as in Meier et al. (1989), or to both as in J. Feagin and Eckberg (1980). With regard to discrimination, it is often not quite clear what is discriminatory about institutional discrimination.

The second and, maybe, biggest issue I have with conceptualizations of institutional discrimination, is that many, if not all, definitions of institutional racism or institutional discrimination equate or conflate discrimination with inequality in some way (see, e.g., Williams, 1985, p. 330; Pincus, 1996). In fact, the idea that institutions, policies, and regulations do *not* discriminate among or against individuals by virtue of a particular characteristic but, nevertheless, lead to inequality between societal groups that can be distinguished according to the characteristic in question is at the heart of the institutional discrimination literature.

Also, many definitions of institutional discrimination or racism unnecessarily restrict the roles particular groups can play in all of this: Then, by definition, only members of or institutions set up by members of "dominant" (Pincus, 1996, p. 186; Sidanius & Pratto, 1999, p. 127; J. Feagin & Eckberg, 1980, pp. 9, 12) groups may discriminate against members of "subordinate" (J. R. Feagin & Booher Feagin, 1986, p. 30; J. Feagin & Eckberg, 1980, p. 12; Sidanius & Pratto, 1999, p. 127) or "minority" (J. R. Feagin & Booher Feagin, 1986, p. 31; Pincus, 1996, p. 186) groups. Now the question is: What characterizes dominant groups and what characterizes subordinate groups? Obviously, the distinction would be redundant if the group that discriminates or set up the institution that discriminates against another group simply denotes the dominant group because of that. If this is not necessarily the case, one and the same behavior could possibly be labeled discrimination if committed by one group but not if committed by another.

Finally, considering the tight connection of early contributions to the institutional racism literature and the US civil rights movement, it is not surprising that many contributions on institutional racism and institutional discrimination implicitly but often also explicitly morally judge and condemn discrimination against particular societal

groups. Take J. R. Feagin and Booher Feagin (1986), for example, whose "discussion assumes that race and sex discrimination are unjust and should be remedied" (J. R. Feagin & Booher Feagin, 1986). Or take the German literature on institutional discrimination that highlights the normative nature of contributions to the institutional discrimination literature and advocates the explicit integration of the concept of institutional discrimination and research on justice and equity (e.g., Gomolla, 2016, p. 18). Particularly noteworthy in a chapter on how to define discrimination are definitions that feature built-in normative judgments such as the one by J. M. Jones (1997) cited above. Such definitions are not useful for empirical social science research, which is why I will certainly not adopt them.

Investigating the institutional and organizational determinants of discrimination—in the sense of causal effects of institutional and organizational level variables on discrimination as defined in section 2.3.3—is certainly of great interest to social science researchers and the public. However, I am not sure that the existence of such effects need to be named in a particular way. Put differently, the term institutional discrimination is certainly not needed to study these effects, even if I do not fully object wholeheartedly. However, given this long list of drawbacks from above, I will not make use of the concept of institutional discrimination in this study. Since it is a widely cited approach that claims to be able to explain discriminatory behavior and between group inequality and, as such, has been applied to German education, I will nevertheless briefly return to it in chapter 3, where I discuss its explanatory power as a theory.

Structural and systemic discrimination

The terms structural racism or structural discrimination and systemic racism or systemic discrimination are muss less used than their institutional counterparts. Also, sometimes, they are used synonymously to institutional racism or discrimination or some variant of it: Pincus (1996, p. 186), for example, defines structural discrimination in a way that equals others' understanding of institutional discrimination and J. R. Feagin and Bennefield (2014, p. 7) introduce systemic racism as synonymous with institutional racism. In the German literature on discrimination in general and discrimination in education in particular, these terms are used rarely and, if so, often interchangeably (Gomolla, 2010; but cf. Gomolla, 2016, where institutionelle and strukturelle Diskriminierung are explicitly distinguished).

I will not go into details on these terms, but merely summarize briefly why I don't see any use for them in this study: Many contributions to this literature do not provide a clear and self-contained definition at all (e.g., Bonilla-Silva, 1997; J. R. Feagin, 2006).

Also, even more so than institutional forms, structural and systemic racism or discrimination are essentially conceptualized as not more or little more than "pervasive racial disparities" (Reskin, 2012, p. 18). Additionally, structural and systemic forms of racism or discrimination are typically conceptualized in a rather narrow way so that they tend to apply to one country—typically the US—and mostly to race relations, not other intergroup relations (e.g., Bonilla-Silva, 1997; J. R. Feagin, 2006; Reskin, 2012). The contributions to this literature usually contain normative language and judgments (especially see J. R. Feagin, 2006; J. R. Feagin & Bennefield, 2014) that I, as laid out in chapter 1, wish to avoid wherever possible.

2.5 Summary and Conclusion

Progress in science comes not from coining new terms. It comes from using established terms in a logically consistent way to explain old and new phenomena by means of making explicit the mechanisms that are at work in bringing these phenomena about. To argue in favor of a logically consistent and useful definition of discrimination is what I have sought to contribute by means of this chapter. I have argued that discrimination is best understood as the causal effect of an information about or a signal sent out by an individual on how this individual is treated by another individual. I have argued that this general definition is the most useful starting point for defining discrimination, since it avoids unnecessary constraints that are hard to justify and since it avoids normative judgments that have to be or are likely to be made along the way of specifying the definition. In the present study, I am concerned with discrimination by teachers based on information about students' ethnicity, social class, and sex or gender, respectively—also known as *ethnic* discrimination, *social class* discrimination and *sex* or *gender* discrimination.

3 Theories of Discrimination

> To be able to measure [...] discrimination of a particular kind, it is necessary to have a theory [...] of how such discrimination might occur and what its effects might be.
>
> *(Blank et al., 2004)*

In this chapter I am concerned with theories, models, or mere hypotheses that claim to be able to help explain discrimination. As with definitions, explanations of discrimination abound. I focus on—but do not limit my discussion to—contributions that have been applied to discrimination in education and that fulfill two general criteria: The first criterion is that I accept as a theory only those contributions that seek to provide a *causal explanation* for discrimination, which requires that a *mechanism* of some sort has to be provided (Elster, 1989). The second, additional criterion, added by methodological individualism (see section 1.5.1), is that the mechanism provided by the theory has to refer to *individual behavior* in some way.

3.1 Economic Theories of Discrimination

Economic theories of discrimination are often said to fall into one of two camps (for this distinction see, e.g., Charles & Guryan, 2011, p. 495; Black, 1995; Kristen, 2006a): Theories in the first camp are built on the seminal work by Becker (1957/1971), who introduced what he called "a taste for discrimination" (Becker, 1957/1971, p. 14). This taste is part of an actor's utility function, often equated with prejudice (e.g., Altonji & Blank, 1999; J. Knowles et al., 2001), sometimes more generally treated as a preference (Guryan & Charles, 2013, p. 418), but virtually always contrasted with information (e.g., Guryan & Charles, 2013; J. Knowles et al., 2001; Levitt, 2004). We will see that in Becker (1957/1971) these distinctions are not perfectly clear.

The second camp is home to theories of statistical discrimination (Aigner & Cain, 1977; Arrow, 1973; Phelps, 1972). These models rest on the notion of imperfect information or information asymmetries. This way, they circumvent the problem of introducing preferences that are external to the market. While it is straightforward to explain individual discrimination using these models, discrimination on the group

level and, thus, inequality cannot—without further qualifications—be explained by all models of statistical discrimination (Aigner & Cain, 1977).

3.1.1 Taste Discrimination

Becker (1957/1971), who provides the first extensive treatment of the economics of discrimination[12], seeks to explain why there are differences in labor market outcomes—mainly wages and employment rates—between various groups of workers in the market against the backdrop of rational actors. To this end, Becker (1957/1971) generalizes conventional theory and extends the utility function of market actors—employers, employees, and consumers—beyond money. He introduces the concept of a *taste for discrimination* that he conceptualizes as follows:

> If an individual has a "taste for discrimination", he must act as if he were willing to pay something, either directly or in the form of a reduced income, to be associated with some persons instead of others. (Becker, 1957/1971, p. 14)

The operational definition of this taste for discrimination is the so called *discrimination coefficient*, often abbreviated DC and denoted d in formal expressions. Becker (1957/1971)'s ingenious and rather uneconomic idea was to distinguish between *money costs* and *net costs* of a transaction and use the DC as a bridge between them (Becker, 1957/1971, p. 14). Formally speaking, instead of merely considering the money wage rate π of an employee, an employer i with DC d_i acts as if $\pi(1 + d_i)$ were the net wage rate.

Mechanics of taste discrimination

Obviously, if $d_i > 0$ toward a particular group, the net costs, $\pi(1+d_i)$, would be higher than the money costs, π, for any transaction with a member of that group. According to Becker (1957/1971), this case is what constitutes discrimination, since

> [d]iscrimination is commonly associated with disutility caused by contact with some individuals. (Becker, 1957/1971, p. 15)

It is through this mechanism that a taste for discrimination turns into differential behavior towards members of different groups. The reverse case of $d_i < 0$ toward a particular group yields lower net costs than money costs and, according to Becker (1957/1971, p. 15), constitutes nepotism.

12 This, "The Economics of Discrimination", is also the title of his book.

Note, however, that Becker (1957/1971) explicitly contradicts Allport (1954) and acknowledges that "the social and economic implications of positive prejudice or nepotism are very similar to those of negative prejudice or discrimination" (Becker, 1957/1971, footnote 3). I agree with Becker (1957/1971), whose position is perfectly compatible with the definition of discrimination as a causal effect proposed in chapter 2 (see section 2.3.1 in particular): Since a causal effect is always defined via the difference of at least two causal states, an advantage for members of group A is equivalent with a disadvantage for members of group B—and vice versa—if only these two groups exist in a given market. If there are more than just two groups, things get a little more complicated since there is not just one but multiple differences between different causal states that have to be considered. Social psychologists, too, have realized that nepotism or *ingroup-favoritism* may be as or even more likely the cause of discrimination as *outgroup-derogation* (Brewer, 1999; Greenwald & Pettigrew, 2014; Tajfel, 1982; Tajfel & Turner, 1986). I will return to these mechanisms and related theories from social psychology below in section 3.3.

Before I show how Becker (1957/1971) explains inequality between groups and before I discuss how to apply his theory to the German education system below, let's have a closer look at the proposed mechanism and the determinants of the DC. Essentially and as stated above, Becker (1957/1971) suggests that rational actors discriminate against individual employees because they seek to maximize their utility. To this end, actors ask for compensation in the form of a wage premium when surrounded by people they dislike, that is, people that negatively contribute to the actors' utility. Alternatively, actors are willing to forfeit income in order to be surrounded by people they like, that is, people that positively contribute to the actors' utility. These contributions to the actors' utility functions are captured by the discrimination coefficient, DC.

Now, what are the determinants of the DC? In fact, Becker (1957/1971) gives several examples for such determinants, including the social and physical distance between an individual and a particular group, their relative socioeconomic status, and the number of members from the group in question (Becker, 1957/1971, p. 16). Interestingly, Becker (1957/1971, pp. 16–17) also explicitly discusses *ignorance* as a determinant of the DC—a fact largely ignored in the literature (but cf. Hunkler, 2014). However, even though Becker (1957/1971) demands that "the amount of knowledge available must be included as a determinant of tastes", his discussion suggests that he takes *prejudice*—that he also uses synonymously with preference—as the major ingredient of the DC:

> Ignorance may be quickly eliminated by the spread of knowledge, while a

> prejudice (i.e., preference) is relatively independent of knowledge (Becker, 1957/1971, p. 16)

Also, neither the DC nor any other parameter in Becker (1957/1971)'s model captures knowledge or ignorance about one single candidate. Put differently, d_i^j only varies between different employers, i, or actors more generally, and target groups, j. That Becker (1957/1971) allows knowledge about groups as a whole to be incomplete but assumes knowledge about individuals to be perfect, strikes me as contradictory: Either the information about individual productivity is perfect, then information about the group should not determine tastes and should, in fact, play no role in the model at all, or information about individuals is more or less imperfect, then, however, we would want to know how to combine knowledge about the group and about the individual to arrive at the net cost of this individual.

However, from the discussion and formal model provided by Becker (1957/1971) it is rather clear that his models mainly rests on prejudice, or, more generally, preferences, and does not acknowledges varying degrees of knowledge about single candidates depending on situational constraints. Note that this is a crucial difference to models of statistical discrimination that explicitly contain a parameter to capture the reliability of information available about a single candidate in a given situation.

Taste Discrimination and inequality

In order to explain group discrimination and, thus, inequality between groups, Becker (1957/1971, p. 17) introduces a *market discrimination coefficient*, abbreviated MDC and defined as

$$MDC \equiv \frac{\pi^A}{\pi^B} - \frac{\pi_0^A}{\pi_0^B} = \frac{Y(A)}{Y(B)} - \frac{Y_0(A)}{Y_0(B)}, \tag{3.1}$$

where π^A and π^B are the observed equilibrium wage rates, while π_0^A and π_0^B are their counterfactuals without discrimination. $Y(A)$ and $Y(B)$ are the actual incomes of A and B, while $Y_0(A)$ and $Y_0(B)$ are incomes without discrimination.[13]

Obviously, the magnitude of the MDC and, thus, the magnitude of wage differentials between A and B, depend on the magnitude of individual DCs (Becker, 1957/1971, p. 18). So, when all employers feature the same taste for discrimination against group B—i.e., when d_i is constant across all i—members of group B either have to accept a wage rate of $\pi(1 - d_i)$ or will not be hired. In such a scenario, the causal effect of

13 Note that Becker (1957/1971)'s notation nicely corresponds to that used in the literature on counterfactual causality and potential outcomes (e.g., Gangl, 2010; Imbens & Rubin, 2015; Morgan & Winship, 2015; Pearl, 2009).

being member of group B instead of A is negative for all in B. Also, all members of B earn less then expected based on their productivity. Inevitably, this aggregates to group discrimination and, thus, inequality.

However, Becker (1957/1971)—who is credited to be the first to realize and theorize this (Heckman, 1998, p. 102)—argues that such a scenario is not realistic and, hence, the assumption that the MDC only depends on individual DCs is mistaken. Indeed, Becker (1957/1971, p. 43) shows that while it is necessary to know the tastes of an average or "representative" employer to assess market discrimination, it is not sufficient. Analyses of discrimination in labor markets, for example, have to take into account processes of self-selection of employees on particular employers. In education processes of self-selection might or might not be of less relevance; for elementary school in Germany they should be less relevant, since in most federal states binding school districts limit school choice considerably. But, also in German education it is not simply the taste of an average teacher that determines discrimination in the education system. See already my discussion in section 2.4.1. I will return to this discussion in chapter 6 in which I present an experiment that gives credit to Becker (1957/1971)'s insights.

Application to the German education system

It has been suggested that not only actors in labor markets but also teachers feature a taste for discrimination and Becker (1957/1971)'s theory has been applied to different educational settings in different countries (e.g., Hanna & Linden, 2012; Kristen, 2006a; van Ewijk, 2011). With regard to tastes or prejudices of German teachers we know very little, indeed. I discuss the most enlightening of the few quantitative studies in chapter 4 before I present my own analyses on teachers' prejudices towards different ethnic groups. What we know and I find confirmed in my analyses is that teachers hold negative prejudices towards Turks. I find that less teachers in Germany hold negative prejudices about Eastern Europeans and virtually none hold negative prejudices against Italians. Just about nothing is known about teachers' attitudes towards different social classes or men and women or boys and girls, respectively.

With regard to the application of Becker (1957/1971), I am more skeptical than others (e.g., Kristen, 2006b): Recall the key mechanism that motivates actors to discriminate, namely "disutility caused by contact with some individuals" (Becker, 1957/1971, p. 15) that actors seek to avoid or demand to be compensated for. However, for both grading and recommending tracks teachers typically cannot alter their own utility by discriminating against students. In contrast to the labor market, where employers

profit when paying lower wages to employees from particular groups or simply not hiring them, teachers may only punish students with bad grades or recommendations for lower tracks. Usually, this should not affect the teacher's utility. Therefore, the prediction derived from Becker (1957/1971) for teacher behavior in these situations should be to expect no discrimination in grading or track recommendations, no matter what d is towards the group in question. Of course, there are situations in which Becker (1957/1971)'s mechanism should result in discrimination: Imagine the teacher groups students according to ability or achievement within tracks and continues to be the teacher for one of the created groups, say the advanced group. Following Becker (1957/1971) we would expect this teacher to group students not in accordance to their true ability or achievement π but in such a way that his or her utility is maximized by excluding some students from groups towards whom he or she has a "taste for discrimination".

3.1.2 Statistical Discrimination

The basic idea of statistical discrimination theory (Aigner & Cain, 1977; Arrow, 1973; Phelps, 1972) is that, in a situation of imperfect knowledge about the true productivity of an employee, rational employers use observable characteristics such as race or sex insofar as they carry information about productivity to estimate the unobserved productivity of an individual employee.

Here, I focus on Aigner and Cain (1977), who provide a review of several models of statistical discrimination including an important critique of Phelps (1972). Some simple formal notation helps to fully appreciate its implications: Since employers lack perfect knowledge about the true productivity, q, of a candidate, they have to rely on an indicator or signal of productivity, y. However, y measures q with error, u:

$$y = q + u \tag{3.2}$$

where $q \sim N(\alpha, \sigma_q)$ and $u \sim N(0, \sigma_u)$. It is assumed that employers know this relation and the distributions of q and u. Therefore, employers know that they can construct a weighted average of observed individual performance, y, and assumed group ability, α, to come up with an estimate, $\hat{q}$:

$$\hat{q} = E(q|y) = (1 - \gamma)\alpha + \gamma y \tag{3.3}$$

where γ is the reliability of the measure, test, or signal.

Explaining individual discrimination using equation 3.3 is straightforward: Let's

say we are interested in discrimination conceptualized as the causal effect of signaling membership in group B instead of A. Following the most simple model by Phelps (1972), employers use equation 3.3 to estimate the productivity of a candidate or an employee from one of two groups, A and B. In this simple model, employers know that $\gamma < 1$ and that average productivity differs between groups so that, e.g., $\alpha^A > \alpha^B$. Then, obviously, candidates that appear to be members of group B will estimated to be of lower productivity than had they appeared to be members of group A. Clearly, this is an account of discrimination as an individual-level causal effect by virtue of membership in group A versus B. This scenario is visualized in the left panel of figure 3.1.

A second, still rather simple model that is also described in Phelps (1972), allows γ to vary between groups. Such a scenario with equal means, $\alpha^A = \alpha^B$, but different reliabilities, $\gamma^A > \gamma^B$, is visualized in the right panel of figure 3.1. If we write γ as

$$\gamma = \frac{Var(q)}{Var(q) + Var(u)}, \tag{3.4}$$

where q is the true ability and u the measurement error in y, there are two ways that lead to $\gamma^A > \gamma^B$. First, groups A and B have ability distributions with same variances, $Var(q^A) = Var(q^B)$, but the test used by the employer measures y less precisely for group B, i.e., $Var(u^A) < Var(u^B)$. That results in $\gamma^A > \gamma^B$, which means that the test has a higher (conditional) reliability for group A than for B. The second case arises from equal error variances, $Var(u^A) = Var(u^B)$, but different (conditional) variances, $Var(q^A) < Var(q^B)$, yielding $\gamma^A > \gamma^B$ again. The consequences can be seen in the right panel of figure 3.1: the slope for A (dashed) is steeper than the one for B (dotted). Individual discrimination by virtue of signaled group membership occurs for all y except where the slopes intersect. However, note that in this scenario the sign of the causal effect changes depending on whether $y > y^*$ or $y < y^*$, where y^* is the point in which the slopes intersect.

What's the evidence?

Evidence in favor of different models of statistical discrimination abounds—evidence is found in many different countries for different sectors of society and different actors using different methodologies (for reviews see, e.g., Altonji & Blank, 1999; Cain, 1986; Charles & Guryan, 2011; Guryan & Charles, 2013). Even the arguably most counterintuitive result, namely that groups with lower γ are favored and, thus, profit from discrimination for lower values of y has been backed up by empirical evidence (Scha-

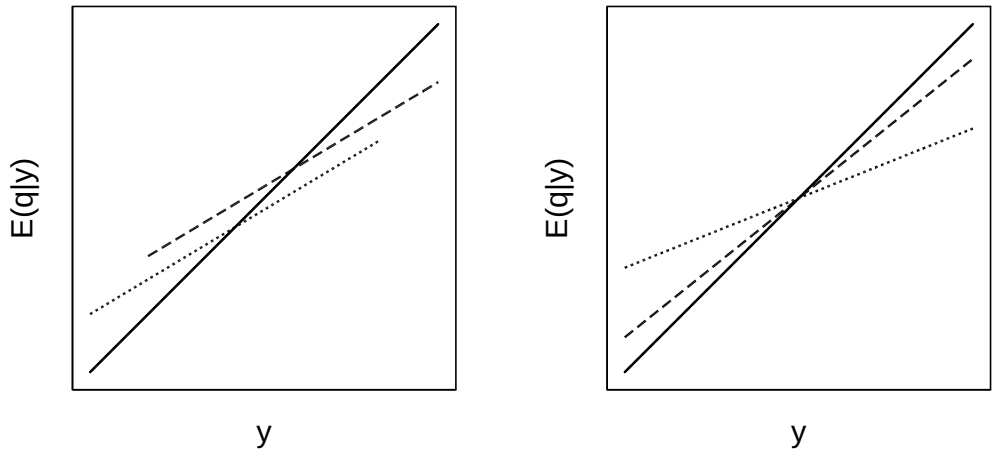


Figure 3.1: Predictions of ability, $\hat{q} = E(q|y)$, by group and test score, y. The bisectrix (solid line) visualizes the case of $\gamma = 1$. The dashed line shows the relationship for group A, the dotted line for group B. The left panel shows two parallel lines indicating $\gamma^A = \gamma^B$, and $\alpha^A > \alpha^B$: $\hat{q}^A = (1 - \gamma^A)\alpha^A + \gamma^A y$. The right panel shows the opposite constellation with equal means, $\alpha^A = \alpha^B$, but different slopes, $\gamma^A > \gamma^B$.

effer et al., 2016). Teachers also seem to act in accordance with statistical discrimination theory (e.g., Hanna & Linden, 2012; van Ewijk, 2011).

With regard to German teachers, we know very little about their general or subject-specific stereotypes about different groups of students—that is, their beliefs about group-specific α—, which is why I examine teachers' stereotypes towards different groups of students in chapter 5. Lorenz et al. (2016), however, find that teachers have systematically lower expectations towards students with a Turkish background, students from families of lower socioeconomic status, and boys, even conditional on relevant controls. This way Lorenz et al. (2016) provide indirect evidence for corresponding stereotypes. My results for more direct measures of teachers' stereotypes in chapter 5 largely confirm the findings of Lorenz et al. (2016).

Evidence on γ comes from studies on accuracy of teacher judgments. The six German studies included in the meta-analysis by Südkamp et al. (2012) find correlation coefficients between teachers' estimates and actual performance of $r = .34$ and $r = .59$, suggesting that teachers are far from perfect in predicting student performance. While teachers overall evaluations might be more precise, note that tasks such as recommending tracks at the end of elementary school involve predictions about future development that might be more more difficult.

Statistical Discrimination and Inequality

As already mentioned in section 2.4.1, models of statistical discrimination have been used to distinguish between individual discrimination and market or group discrimination, respectively, without referring to market forces. In fact, the discussion by Aigner and Cain (1977) starts off as a critique of the model by Phelps (1972), that, according to Aigner and Cain (1977), is not able to explain inequality between groups as a consequence of discrimination. Put differently, the mechanism Phelps (1972) suggests, only accounts for individual discrimination, not group discrimination.

Recall the two simple models from above, visualized in figure 3.1. If the assumptions hold that $q \sim N(\alpha, \sigma_q)$ and $u \sim N(0, \sigma_u)$ and that employers or teachers, for that matter, know these distributions, neither model can explain discrimination on the group level, if interest lies in *average effects*, that is in $E(\hat{q})$, on the group level. This is because for both models Aigner and Cain (1977) show that, on average, employers or teachers, are getting it right, so that $E(\hat{q}^A) = \alpha^A$, $E(\hat{q}^B) = \alpha^B$, $E(q^A) = \alpha^A$, and $E(q^B) = \alpha^B$. Group discrimination, expressed as a difference-in-difference, then, equals zero:

$$\begin{aligned} GD^{A-B} &= (E(\hat{q}^A) - E(q^A)) - (E(\hat{q}^B) - E(q^B)) \\ &= (\alpha^A - \alpha^A) - (\alpha^B - \alpha^B) \\ &= 0 \end{aligned}$$

That is, as long as teachers' beliefs about group means, α, are correct and the test or teachers' perception of individual performance, y, is not systematically biased, individual discrimination does not aggregate to group discrimination.

Now, under which conditions does statistical discrimination lead to group discrimination and, thus, helps to explain inequality? There are several scenarios under which the statistical discrimination mechanism from equation 3.3 leads to group discrimination. One such scenario that Aigner and Cain (1977) discuss, features risk averse decision makers, who seek to minimize the risk of underestimating the ability of workers or students. As a consequence employers or teachers apply a higher risk penalty to groups with lower γ. This may result in group differences in wages or track recommendations conditional on ability and, thus, group discrimination. There is indeed evidence that teachers in Germany are risk-averse when recommending tracks: They tend to recommend the lower out of two tracks in ambiguous situations—when, for instance, children have a grade point average in between two cutpoints (Maaz et al., 2008).

Another class of scenarios in which individual discrimination aggregates to group

discrimination gives up the assumption of being interested in means only and the of normally distributed outcome variables only. An important case in point for research on discrimination in the German education system is that the outcome of interest might not be normally distributed but a categorical variable, such as school tracks. If interest, for example, lies in the highest track to which students may only go when $y > y^*$, where y^* is a cutoff point somewhere in the distribution of y, then members of group B will suffer from group discrimination when $\gamma^A > \gamma^B$ as in the right panel of figure 3.1.

Group discrimination might also arise if the key assumption of correct beliefs about α, that is, correct stereotypes, is violated. Thus, in case teachers' stereotypes are biased to the disadvantage of a group, this group would suffer from group discrimination. England and Lewin (1989) explicitly discuss this model that they call "error discrimination". Bohren et al. (2019) refer to the same phenomenon as "inaccurate statistical discrimination" and show that ignoring the possibility of incorrect stereotypes may lead to a confusion of inaccurate statistical discrimination with other forms of discrimination, such as taste discrimination. I will provide evidence for biases in teachers' stereotypes in chapter 5.

Application to the German education system

As with taste discrimination, the question is whether or not the mechanisms suggested by models of statistical discrimination can be meaningfully applied to German education and which general hypotheses can be derived from applying them.

With regard to track recommendations at the end of elementary school in Germany, Kristen (2006b) and others (e.g., T. Schneider, 2011) suggest that, following models of statistical discrimination, we should not expect discrimination—i.e., neither individual nor group discrimination—to occur, since teachers would possess perfect or almost perfect knowledge about individual students after teaching them for years. However, teachers face the task of predicting future performance at different tracks they typically will not know in great detail so that I see room for statistical discrimination to slip in teachers' track recommendations[14]. Concerning grades, especially grades for written exams, I would expect little to no discrimination, as observed performance should typically provide the teacher with all relevant information and, thus, γ should be close to one.

Note that statistical discrimination—and the contrast with taste discrimination—is

14 I have made this point earlier in my diploma thesis (Wenz, 2009). It is also made in Diehl and Fick (2016).

a great example for how micro-mechanisms matter, even when the interest of the researcher only lies on institutional-level variables. As I will discuss in some more detail below in section 2.4.7, contributions to the institutional discrimination literature have not only failed in providing but also in merely explicitly using a mechanism of human behavior as a microfoundation. Imagine we would be interested in the question of whether or not it makes a difference for the degree of discrimination by virtue of ethnic or social background of the student, when elementary school lasts longer. What we would need to know is not only how the conditions change under which teachers make their decisions, that is, in which ways the logic of the *situation* changes, but also how, in general, teachers act or behave, for that matter, that is, we need to know the logic of *selection*.

Concerning the logic of the situation, the institutional change of tracking students two years later would certainly mean that students are older and, thus, more developed with regard to cognitive and non-cognitive skills. The institutional change might also mean that teachers have taught the same students for a longer period of time and, thus, know them better, but teacher turnover might interfere with this consequence.

Now, for the hypothesis what this institutional change means, it makes a crucial difference whether teachers behave in accordance with Becker (1957/1971)'s model of taste discrimination or in accordance with a model of statistical discrimination such as given in equation 3.3. If only tastes determine discriminatory judgments and behavior, the degree of discrimination against any group should not be affected by such an institutional change. However, when students are older or when teachers know their students for a longer time, teachers might perceive observed student performance as more reliable, so that γ increases. Following equation 3.3 we would expect that the institutional change leads to less individual discrimination in teachers' track recommendations and, since the outcome is categorical, also to less group discrimination.

3.2 Sociological Theories of Discrimination

Sociological contributions to the study of prejudice, discrimination, and intergroup relations more generally have a long history (e.g., Blalock, 1967; Blumer, 1958; Bobo, 1999; Bobo & Hutchings, 1996; Bogardus, 1925, 1933, 1958; Myrdal, 1944; R. E. Park, 1924; Quillian, 1995, 2006; Sumner, 1906). However, the only theoretical perspective that has been applied to education repeatedly in both international and German literature, seems to be the perspective of institutional racism or institutional discrimination, respectively.

3.2.1 Institutional, Structural, and Systemic Discrimination

At the end of chapter 2, I have discussed several definitions of institutional racism and institutional discrimination. Here, I will briefly comment on institutional, structural, and systemic discrimination as theories. As in section 2.4.7, I focus on the literature on institutional racism and discrimination. The problems of structural and systemic racism and discrimination approaches are virtually the same—if anything, they are greater.

As a theory, I am even more critical towards contributions to the institutional racism and institutional discrimination literature than towards the attempts to define the phenomenon. In fact, I think that it is justified to say that there really is no institutional racism or institutional discrimination *theory*—not to mention *a* theory. I think so, since the contributions to this literature have not managed to provide a clear and comprehensible mechanism that actually helps to explain how institutions—conceptualized in whatever way—affect discrimination or at least disparities between groups. This lack of explanatory power has even been recognized from advocates of institutional racism and discrimination (e.g., Gomolla, 2016; Troyna & Williams, 2012; Williams, 1985). However, little progress has been made towards overcoming this gap (Gomolla, 2016, p. 7).

While I second the critique of those who demand a microfoundation of institutional discrimination approaches (e.g., by incorporating social psychological mechanisms; Berard, 2008), note that there are others who criticize the institutional discrimination literature for quite the contrary, namely for falling back on micro-mechanisms and falling short of providing more detailed descriptions of macro-level or institutional-level mechanisms (e.g., Troyna & Williams, 2012; Wight, 2003). All this is not to say that nowhere in this literature can be found suggestions on mechanisms in general and mechanisms of individual human behavior in particular that might carry a causal effect of institutions on discriminatory individual behavior. Remarkably though, such ideas can be found more often in earlier contributions: J. R. Feagin and Booher Feagin (1986), for instance, wrote:

> [...] whatever the scale of the organizational context all discrimination involves individual actors. The "bottom line" in all types of discrimination is someone actually doing something to someone else. Large corporations and bureaucracies *do not act* except in some metaphorical sense; the people in them do act, even though they may be routinely carrying out required regulations inherited from some dusty past. (J. R. Feagin & Booher Feagin, 1986, p. 25, their emphasis)

While this short paragraph can certainly be read as a plea for a microfoundation of institutional discrimination and while the reference to routine behavior might have been a helpful starting point, J. R. Feagin and Booher Feagin (1986) explicitly provide neither a microfoundation, nor a mechanism more generally. Unfortunately, later contributions to the institutional discrimination literature have moved even more towards structural and systemic approaches and away from individual mechanisms (e.g., Bonilla-Silva, 1997; J. R. Feagin, 2006; J. R. Feagin & Bennefield, 2014).

As for an application to the German education system or really any education system, the study of causal effects of institutions on discrimination by virtue of characteristics such as race or ethnicity, social class, and sex or gender is certainly a relevant one. It is of interest to the scientific community as it allows indirect tests of different theories of discrimination including their micro-mechanisms. And it should be of interest to a broader audience as it carries policy implications, if, for example, it can be shown that the magnitude of ethnic or social class discrimination in teachers' recommendations at the end of elementary school can be reduced by tracking students at a later age, when teachers can predict more precisely the students' development in the coming years.

3.3 Social Psychological Theories of Discrimination

From single hypotheses over theoretical models to more complex theories, the discipline of social psychology has produced more evidence on stereotypes, prejudice, and discrimination than any other. However, a closer look reveals that many findings from social psychological studies from the last decades have focused more on stereotypes and prejudice rather than on discrimination. In this section I discuss three of the most important approaches and models that I deem useful to explain discrimination and applicable to education. Also, the theories and models discussed in the present section are among the most widely cited and applied in recent decades.

3.3.1 Social Identity Theory

It is no recent observation (e.g., Allport, 1954; Sumner, 1906) that humans tend to hold negative stereotypes and prejudices about outgroup members, hold positive stereotypes and prejudices about ingroup members, and discriminate among people by virtue of the distinction between ingroup and outgroup. One of the most prominent social psychological theories that tries to explain why this is, is social identity theory (SIT; Tajfel, 1982; Tajfel & Turner, 1986). It provides a clear micro-mechanism for why

stereotypes, prejudices, and discriminatory behavior should be biased in favor of ingroups and ingroup members.

SIT: Its mechanics and evidence

Its basic assumptions are the following (Tajfel & Turner, 1986, p. 16): Humans have a need for self esteem and a positive self-concept. Self-esteem and self-concept, in turn, depend on both personal and social identity, which is why humans strive for positive personal and social identities. While the personal identity influences self-esteem via evaluations of personal achievements, the social identity or identities of a person can influence self-esteem via the evaluations of groups we do or think or feel we belong to. Key to understanding how social identity theory explains biased stereotypes, prejudice, and discrimination is that what counts for a positive self-concept is the relative evaluation of groups with reference to other groups. Thus, there are two mechanisms that provide an alternative route to affect self-esteem: Ingroup-favoritism and outgroup-derogation. Hence, one major prediction of social identity theory is that we tend to think better of members of our own group and our ingroup as a whole and derogate members of groups we do not belong to, that is, outgroups, in order to achieve higher self-esteem or compensate for low self-esteem (Fein & Spencer, 1997). We might think of people from our own group as more sympathetic and smart and of people from other groups as unappealing and stupid.

The key hypothesis of social identity theory is the *self-esteem hypothesis*, that can be split into two parts: First, behavior through which a person favors ingroups or derogates outgroups as a whole, or respective group members, should enhance a person's self-esteem. Secondly, the higher the need for self-esteem, the higher the likelihood that a person engages in discriminatory behavior that favors ingroups or derogates outgroups. Fein and Spencer (1997)'s classic study, for example, provides evidence for both mechanisms: Participants whose self-esteem had been lowered by negative feedback evaluated a woman more negatively when she was (supposedly) Jewish than when she was (supposedly) Italian. Those among the negative-feedback candidates given the opportunity to belittle the Jewish woman showed a post-experiment increase in self-esteem.

By and large, literature reviews (Abrams & Hogg, 1988; M. Rubin & Hewstone, 1998) suggest that the evidence of numerous empirical studies is in favor of social identity theory and the self-esteem hypothesis. However, the self-esteem hypothesis seem to be more applicable to "specific, social, and state forms of self-esteem than to global, personal, and trait forms" (M. Rubin & Hewstone, 1998, p. 50). With regard to ingroup

favoritism versus outgroup derogation, the evidence suggests that stereotypes, prejudice, and discrimination are mainly motivated "by the desire to promote and maintain positive relationships within the ingroup rather than by any direct antagonism toward outgroups" (Brewer, 1999). Recall, however, that when ingroup and outgroup serve as groups of comparison, that is, the potential outcomes of being treated as an ingroup member compared to being treated as an outgroup member are of interest, the difference does not matter.

Application to the German education system

Applying social identity theory to the situation of teachers in German education is straightforward. Recall that actual group achievements contribute to a person's social identity, which, in turn, satisfies the need for self esteem. The crucial point about the situation at the end of elementary school in Germany is that it enables teachers to actually influence the educational achievement of different groups of students. Take Turkish students, for example, that are outgroup members for teachers with a German background. Besides the possibility of favoring students of German background over students with Turkish background by holding more positive stereotypes and prejudices about the Germans compared to the Turks, teachers may favor German students over Turkish students when grading exams or recommending a secondary school track. For the difference between these groups of students it is, obviously, irrelevant whether this pattern arises due to students of Turkish origin receiving lower grades or recommendations than they deserve or students without immigrant background receiving higher grades or recommendations than they deserve.

In conclusion, social identity theory (SIT) seems to be a useful complement to theories such as Becker (1957/1971)'s theory of taste discrimination or theories of statistical discrimination (e.g., Aigner & Cain, 1977) because it helps to explain preferences against particular (out-)groups or stereotypic beliefs about a group's mean ability level or other characteristics. Last but not least and in addition to being a complement to other theories, SIT is a powerful alternative—in particular to Becker (1957/1971)'s model of taste discrimination—as it can be used to derive predictions about discrimination in grading or tracking more directly, as I have shown in this section.

3.3.2 The Continuum Model

Dual process models were invented by social psychologists and cognitive psychologists in the 1980s to account for seemingly contradictory or inconclusive findings in empir-

ical research on stereotypes, prejudice, and discrimination that more simple models or mechanisms had trouble explaining (see, e.g., Gawronski & Creighton, 2013, for a review). For a study on discrimination, maybe the most important motivation for the development of dual process models was the question of how to explain the moderate correlation of attitudes and behavior:

> By shifting the focus from asking *"Do attitudes guide behavior?"* to the question, *"How do attitudes guide behavior?"* dual process theorizing provided important insights into the conditions under which attitudes do or do not influence behavior. (Gawronski & Creighton, 2013, p. 286)

The continuum model: Its mechanics

Here, I focus on the *continuum model* (Fiske et al., 1999; Fiske & Neuberg, 1990), one of the earlier and rather popular models that is also rather general and covers affective, cognitive, and behavioral outcomes—that is, prejudice, stereotypes, and discrimination, respectively. The continuum model starts from the observation that automatic and immediate categorization of others is a general and basically inevitable process of social cognition that enables individuals to quickly distinguish between ingroup and outgroup members (Banaji & Hardin, 1996; Zarate & Smith, 1990) . Thus, it suggests categorization as default cognitive process in person perception. To explain and predict when people engage in the default process of a category-based response and when in a piecemeal-based response, the model relies on "two primary factors: the available information and the perceiver's motivation" (Fiske et al., 1999, p. 232).

Only if the target is of minimal interest or relevance for the perceiver in the very moment of categorization, perceivers are motivated to allocate attention to individuating information and move down the continuum from category-based judgments toward a "piecemeal integration" (Fiske et al., 1999, p. 233) of individual attributes. This process of recategorization and, eventually, piecemeal integration may only be started if the available information is rich enough and the perceiver has the time and the cognitive capacity to take it into account. Put differently, only if motivation is high and information rich enough, can we expect that discrimination on the basis of prejudices and stereotypes does not occur.

What's the evidence?

Overall, social psychological dual process theories, including the continuum model, have received support from numerous empirical studies (Gawronski & Creighton,

2013, p. 294). In particular, both Fiske and Neuberg (1990) and Fiske et al. (1999) provide plenty of evidence for the core premises of the model including the role of information available and motivation to overcome simple category-based responses. However, the assumption that category-based responses are the default over all situations has been challenged (Chun & Kruglanski, 2006).

Application to the German education system

While teachers may be motivated to overcome category-based judgment when dealing with their students, the information available might not always suffice depending on the situation. When grading a manifest performance, for example, sufficiently motivated teachers should not show any discriminatory biases. However, when the same teachers need to predict future development of students when recommending tracks at the end of elementary school, the information at hand might not be rich enough to move down the continuum all the way to a "fully individuating impression" (Fiske & Neuberg, 1990, p. 1). Thus, it might be necessary for the teacher to combine individuating characteristics with category-based information such as stereotypical beliefs. In such a situation, the continuum model would predict a causal effect of the teachers' stereotype on the behavior towards the student and, thus, discrimination by virtue of the category the student was assigned to.

3.3.3 Aversive Racism

Gaertner and Dovidio (1986) developed the aversive racism approach to explain why discrimination against blacks in the US continued even though support for openly racist stereotypes, prejudices, and policies had been in decline for many years. While other approaches—including symbolic racism (Sears & Henry, 2005), modern racism theory (McConahay, 1983, 1986), and ambivalent sexism (Glick & Fiske, 1996, 2001)—were developed with similar aims (for concise reviews see, e.g, Dovidio et al., 2017; Nier & Gaertner, 2012), I chose to discuss aversive racism over these other approaches since these alternatives are concerned much more with the content and valence of prejudice and stereotypes towards racial minorities such as blacks, and women. They are less concerned with explaining discrimination, which might be the reason for why they lack explicit mechanisms of human behavior. Put differently, it largely remains unclear under which conditions which beliefs or attitudes are overtly expressed or acted out.

Mechanics of Aversive Racism

The aversive racism approach is built on the idea that while explicit and blatant stereotypes and prejudice against blacks and other minorities might have declined, implicit and more subtle beliefs and attitudes might still be held by many if not all members of the white majority in the US. In fact, the theory explicitly targets the beliefs, attitudes, and behavior of so-called "aversive racists", who

> […] sympathize with the victims of past injustice; support public policies that, in principle, promote racial equality and ameliorate the consequences of racism; identify more generally with a liberal political agenda; regard themselves as nonprejudiced and nondiscriminatory; but, almost unavoidably, possess negative feelings and beliefs about blacks. (Gaertner & Dovidio, 1986, p. 62)

That is, instead of a "taste for discrimination" (Becker, 1957/1971) or otherwise consistent attitudes that favor ingroups over outgroups, individuals may hold rather inconsistent and ambivalent attitudes. Now, Dovidio and Gaertner (2008), Gaertner and Dovidio (1986) reckon that, for understanding and predicting behavior, "[o]ne key element is the nature of the situation" (Dovidio & Gaertner, 2008, p. 45).

The other key element, of course, is a mechanism: Aversive racism theory suggests that individuals are motivated to sustain a positive self-image, so that, in situations where the corresponding social norms are salient and behavior is overt and identifiable, individuals who explicitly endorse egalitarian values and see themselves as nonprejudiced, would *not* discriminate. In situations, however, in which the corresponding norms are not salient enough to trigger the explicit egalitarian attitudes, in which behavior can be acted out more covertly, or in which discriminatory behavior to the disadvantage of, say, Blacks, can be rationalized on the basis of another factor than race, the same individuals would, indeed, discriminate on the basis of race and, thus, follow their implicit and unconscious racist attitudes (Dovidio & Gaertner, 2008, pp. 45–46).

How useful is Aversive Racism as a theory?

Obviously, the starting point of aversive racism theory is rooted in phenomena and observations at a particular place and time involving particular groups in specific roles—some but not all Whites as "aversive racists", Blacks as victims. However, the general mechanism—i.e., the motivation to uphold a positive self-image as unprejudiced nondiscriminator—is applicable to other places, times, and groups of people. Since it

clearly focuses on individual behavior, aversive racism theory satisfies the criteria I put up in the beginning of this chapter.

But what's the evidence that people actually work this way? With regard to the more general mechanics, there is plenty of evidence in favor of aversive racism theory: First, it is well documented that implicit measures of beliefs and attitudes do not perfectly coincide with explicitly reported beliefs and attitudes (Cameron et al., 2012; Devine, 1989; Dovidio et al., 2002; Greenwald et al., 2009; Hofmann et al., 2005; Nosek, 2007). Secondly, we know that people are indeed motivated to maintain a positive self-image (e.g., Baumeister, 1982; Baumeister et al., 1989)—see also the contributions to the literature on social identity theory above in section 3.3.1.

I have also introduced aversive racism as an explanation for discrimination in education, since its key mechanisms seem to be perfectly compatible with more or less wide rational choice models of human behavior (see, e.g., Kroneberg & Kalter, 2012). That individuals do not openly express negative stereotypes and prejudices and do not engage in discriminatory behavior in an "era of contested prejudice" (Lucas, 2008), can certainly be understood as sanction-avoiding behavior and, thus, behavior that maximizes subjective expected utility. But even without external sanctions, in a wider rational choice model, internalized egalitarian and liberal norms could be expected to lead to the behavior predicted by aversive racism theory. Also, in later publications (e.g., Dovidio & Gaertner, 2008), the model was explicitly linked to the distinction between implicit and explicit cognition and, therefore, might very well be integrated into general dual process models of action, such as the model of frame selection, a recent sociological contribution that integrates rational calculating behavior with automatic spontaneous behavior (Esser, 2001; Kroneberg, 2010; Kroneberg & Kalter, 2012; Kroneberg et al., 2010).

Application to the German education system

Conditions and mechanisms suggested by aversive racism theory are readily applied to the German education system and German teachers. With regard to the situations of interest in this study—i.e., grading and track recommendations—I have already argued in this chapter (see, mainly, section 3.1.2) that both the grade for a single assignment as well as a final grade leave room for interpretation, especially in German elementary school, where standardized testing and grading are rare. In many states teachers also have a fair amount of leeway when recommending secondary tracks at the end of elementary school, but regulations differ and in several states recommen-

dations depend more or less perfectly on grades that, themselves, however, are given by teachers.

I have also argued above in this chapter that teachers in German schools, while certainly holding explicit negative stereotypes and prejudices, probably hold less negative explicit prejudices against students with a Turkish background as well as against students with a lower social class background than the general public. Since we do not know much about it, I take a closer look at the explicit attitudes of teachers in Germany in chapter 4. However, a hint on what to expect comes from Hachfeld et al. (2011), who report the results of a study with teacher candidates and educational science students, who turn out to score low on explicit measures of prejudice but high on explicit measures of both multicultural and egalitarian beliefs (Hachfeld et al., 2011, p. 992).

That, at the same time, teachers in Germany hold negative implicit attitudes about certain groups also seems plausible: First, that people hold negative implicit attitudes about outgroups in general and racial or ethnic minorities in particular or at least implicitly prefer ingroup over outgroup members is a well documented global phenomenon (see, e.g., Axt et al., 2014; Cunningham et al., 2001; Greenwald & Banaji, 1995; Greenwald et al., 1998; Nosek et al., 2007). Secondly, Glock and Karbach (2015) report the results of an experimental study, in which German preservice teachers showed implicit preferences of ethnic majority students over ethnic minority students, based, in part, on negative implicit attitudes towards ethnic minority students.

3.4 Summary and Conclusion

In this chapter, I have discussed several popular theories of discrimination. I have started with two theories from economics, Becker (1957/1971)'s theory of taste discrimination and statistical discrimination (Aigner & Cain, 1977; Arrow, 1973; Phelps, 1972). I have argued that, while there is evidence that both are important theories to understand discrimination in many different contexts, the mechanism underyling Becker (1957/1971)'s theory is barely applicable to two key situations in German education, namely grading and recommending tracks at the end of elementary school. In contrast, I deem statistical discrimination models and the proposed mechanism applicable and more helpful in potentially explaining discrimination than others (e.g., Kristen, 2006b; T. Schneider, 2011). It is applicable to both grading situations and track recommendations. Also, several models that built on the statistical discrimination mechanism may not only explain discrimination on the individual level but also on the group level and, thus, inequality between groups.

From the sociological contributions to discrimination, I have focused on the institutional discrimination perspective since it has repeatedly been applied in international and German studies on discrimination in education. However, I am rather skeptical that institutional discrimination provides us with an enlightening perspective on discrimination in education as the contributions to this literature lack both theoretical mechanisms in general and a microfoundation in particular. Remarkably, this has been recognized as a problem but not properly addressed in the literature on institutional racism and discrimination (e.g., Gomolla, 2016; Troyna & Williams, 2012; Williams, 1985).

More fruitful are models from social psychology. I discussed social identity theory (SIT; Tajfel, 1982; Tajfel & Turner, 1986), the continuum model (Fiske et al., 1999; Fiske & Neuberg, 1990), and aversive racism (Dovidio & Gaertner, 2008; Gaertner & Dovidio, 1986). These models are all useful theoretical models and readily applied to education, since they offer micro mechanisms that causally explain why discrimination by virtue characteristics such as ethnicity, social class, or gender should occur. However, the models rely on quite different mechanisms that, applied to education, should lead to rather different predictions about discrimination in different situations such as grading or recommending tracks. While SIT relies on a mechanism linking the social identity of a person and, thus, membership in social groups with individual well being and needs, both continuum model and aversive racism explicitly theorize the role of imperfect information and situational ambiguity for the likelihood of engaging in overt discrimination. Last, but not least, the models from social psychology do not explicitly distinguish and, thus, do not explicitly theorize the distinction between individual discrimination and group discrimination. Because all at least incorporate prejudice or ingroup-favoritism in some way—even though moderated by situational influences in case of the continuum model and aversive racism—they are potentially able to explain group discrimination.

I return to these theories and their predictions in particular in chapter 6, when I present an experiment I conducted to examine discrimination by teachers in different situations. Taken together, the theories discussed in the present chapter highlight the importance of the two major determinants of discriminatory behavior, namely prejudice and stereotypes. I examine the prejudices of German teachers towards different groups in chapter 4 and their stereotypes towards different groups of students in chapter 5.

4 Prejudices of German Teachers

> Two words, to the wise researcher, should be sufficient: Study prejudice.
>
> *(Fiske, 1998)*

Prejudice has been conceptualized as or related to attitudes (e.g., Blalock, 1967; Brown, 2010; Ehrlich, 1973; J. M. Jones, 1997; Simpson & Yinger, 1972), evaluations (Correll et al., 2010), emotions (Brown, 2010; Simpson & Yinger, 1972), affections (Correll et al., 2010; D. J. Schneider, 2004), or preferences and tastes (Becker, 1957/1971). In what is known as tripartite model of attitudes or, more generally, tripartite perspective on category-based reactions towards groups or individuals, prejudice is typically described as the affective component, while stereotypes are seen as the cognitive component, and discrimination as the behavioral component (Correll et al., 2010; Fiske, 1998; Zanna & Rempel, 1988).

Studying prejudice in more detail in a study of discrimination in education has both theoretical and empirical reasons. *Theoretically*, prejudice is a major determinant of discrimination and, as discussed in section 2.4.1 and chapter 3, plays a key role in explaining both individual *and* group discrimination and, hence, inequality between groups. Recall from the discussion in chapter 3 that a major difference to stereotypes and beliefs is that, theoretically, prejudice is expected to lead to discrimination on both individual and group level when applied or acted out towards individuals.

As it turns out, prejudice is indeed the better *empirical predictor* of discriminatory behavior than stereotypes and it seems that there is no better predictor of discrimination by virtue of variables such as race, sex, or class than prejudice—except the intention to discriminate. In a meta-analysis of 53 studies, published between 1930 and 1993, Schütz and Six (1996) investigate 60 effect sizes, reporting an average correlation of $r = .37$ between prejudice and intended discrimination, of $r = .49$ between intended discrimination and actual discriminatory behavior, and of $r = .29$ between prejudice and actual discriminatory behavior. Although Schütz and Six (1996) do not explicitly compare the predictive power of prejudice with other constructs such as stereotypes, they conclude that "all of these are less useful than prejudice" (Schütz & Six, 1996, p. 457). Such a comparison was undertaken by Talaska et al. (2008), who report 136 effect sizes from 57 studies that appeared in 54 publications from 1950 to 2002. The authors report the highest median correlations with discriminatory behavior for measures of behavioral intentions ($r = .39$), emotions and emotional preju-

dice ($r = .35$), and combinations of overall valence and emotion ($r = .32$). While the median correlation of discrimination with stereotypes is reported to be $r = .26$, correlations with other belief- and stereotype-related measures are reported to range from $r = .24$ to $r = .08$. In a regression model controlling for possible confounders of the effect size differences between attitudinal measures and measures of beliefs, Talaska et al. (2008) find support for the claim that prejudice is a stronger predictor of discrimination than stereotypes. In fact, on average and controlling for relevant covariates, the correlation of prejudice and discrimination is $\beta = .32$ units higher than the correlation of stereotypes and discrimination (Talaska et al., 2008, p. 282). Taken together, theoretical and empirical reasons substantiate Fiske (1998, p. 373)'s invitation to researchers in intergroup relations: "Study prejudice."

Therefore, in this chapter, I investigate teachers' prejudices towards different groups of students. As throughout this dissertation, my main interest lies with the students of Turkish origin. But, for the sake of comparison and in its own right, I am also interested in other ethnic groups as well as students of different social classes and of different gender.

4.1 Conceptualizing Prejudice

In this chapter and throughout this dissertation I conceptualize *prejudice* as *an attitude toward a particular group or category of people* (see also, e.g., Correll et al., 2010; Ehrlich, 1973; J. M. Jones, 1997; D. J. Schneider, 2004). Since I understand attitudes as "general evaluations of people, objects, and issues" (Fazio & Petty, 2008, p. 1), prejudice is simply an evaluation of a group or category of people. Especially older, traditional definitions of prejudice are less general. In the remainder of this section, I will briefly discuss some of these older but also some more recent definitions. I focus on problems that make—especially the older definitions—much less useful for empirical research than the definition I chose. However, I also present definitions that I largely agree with and, thus, built on.

4.1.1 Less Useful Perspectives on Prejudice

One of the first and certainly the most widely cited definition of prejudice is the one by Allport (1954):

> Ethnic prejudice is an antipathy based upon a faulty and inflexible generalization. It may be felt or expressed. It may be directed toward a group

> as a whole, or toward an individual because he is a member of that group. (Allport, 1954, p. 9)

I have several issues with Allport (1954)'s definition (see Brown, 2010, for a similar critique). First, defining prejudice as an antipathy rules out that there are positive or sympathetic prejudices. However, defined as an attitude, "logically, prejudice can take both positive and negative forms" (Brown, 2010, p. 4). Also, in light of my general definition of discrimination in chapter 2 and the insights from different theories of discrimination such as Becker (1957/1971)'s taste discrimination or mechanisms such as ingroup-favoritism (Tajfel & Turner, 1986; Brewer, 1999), it is evident that positive evaluations of some groups—but not others—are, in effect, just as problematic as negative evaluations. Put differently, if all groups are evaluated positively, but some are evaluated more positively than others, members of groups that are evaluated more positively than others might receive preferable treatment only because they are member of the positively evaluated group, which constitutes discrimination under almost any definition out there, including the one I proposed in chapter 2.

Thus, it seems wise to not restrict prejudice to be negative by definition, but to allow for both negative and positive prejudice to exist in principle. Unfortunately, conceptualizing prejudice as having negative valence has been very common (e.g., Levin & Levin, 1982, p. 65; J. M. Jones, 1972, pp. 2–4; Fishbein, 2002, pp. 4–5; American Psychological Association, 2006). Surprisingly, Brown (2010), too, holds on to this perspective—even though he tries to water it down a little bit, by referring to prejudice as an attitude "which directly or indirectly implies some negativity or antipathy towards that group" (Brown, 2010, p. 7).

My second issue with Allport (1954)'s definition is that it conceptualizes prejudice as faulty or based on something faulty. I join Brown (2010) in rejecting such a restriction mainly because it implies that the correctness of prejudice or its foundation could be assessed. However, defined as an attitude, it can itself neither be true nor false as it does not contain—in contrast to stereotypes or beliefs (see chapter 5)—factual or empirical statements. Of course, prejudices and, thus, evaluations of groups are—at least in part—built on stereotypes and beliefs (Crandall et al., 2011) that themselves might very well be false but could also be pretty accurate. Also, negative prejudices can be built on accurate stereotypes and vice versa. Note that the idea that stereotypes and prejudices as well as discrimination are interrelated concepts is held by many and sometimes called the tripartite model of category-based responses or attitudes (Fiske, 1998, pp. 357, 372; Correll et al., 2010, pp. 45–46; Cuddy et al., 2007), to which I return briefly below.

Last, not least, Allport (1954)'s definition implies that prejudices are inflexible, hard

to change, constructs. While this might be the case empirically, prejudices should not be conceptualized this way. Otherwise empirical research on this question is either ruled out or findings of not so hard to change attitudes would be evidence that the attitude in question would not be a prejudice. Also, what does hard to change or inflexible, for that matter, mean, anyway? Note that this qualification, too, has been picked up by others and used to define prejudice (e.g., Simpson & Yinger, 1972, p. 24).

4.1.2 More Useful Perspectives on Prejudice

Especially recent but also some older definitions of prejudice are more useful than Allport (1954)'s and others' definitions that I have criticized above. Take, for example the definition by Ehrlich (1973) that is virtually the same as the one I proposed above:

> Prejudice can then be defined as an attitude toward any group of people. (Ehrlich, 1973, p. 8)

However, since Ehrlich (1973) conceptualizes attitudes more broadly and, essentially, in a tripartite manner with "cognitive, behavioral, and affective dimensions" (Ehrlich, 1973, p. 4), his understanding of prejudice seems to be broader than mine. Other general conceptualizations of prejudice include the widely cited definition of J. M. Jones (1997):

> Prejudice is a positive or negative attitude, judgment, or feeling about a person that is generalized from attitudes or beliefs held about the group to which the person belongs. (J. M. Jones, 1997, p. 10)

Note that, in contrast to J. M. Jones (1972), J. M. Jones (1997) allows prejudice to be both negative and positive.

Before contrasting prejudice with related constructs in the next section, I should say that I conceptualize prejudice as *individual* attitude, not any form of socially shared attitude (Brown, 2010, pp. 8–11). Of course, in no way does this rule out societal forces determining individual attitudes and, thus, prejudices (e.g., Blumer, 1958; Bobo, 1999; Bobo & Fox, 2003; Crandall & Stangor, 2005; Quillian, 1995). In fact, following methodological individualism as outlined in section 1.5.1, both the determinants and consequences of individually held prejudices may lie on the societal level.

4.1.3 Prejudice and Related Constructs

Understood as an attitude, prejudices are evaluations. In contrast, stereotypes, understood as beliefs, lack any evaluative component. Thus, the key difference is that,

while stereotypes can be correct or more or less incorrect, prejudices cannot. An evaluation is positive or negative but neither false or inaccurate nor true or accurate, respectively. However, stereotypes—sometimes referred to as the cognitive component of prejudice (Dovidio et al., 2010, p. 5; Fiske, 1998, p. 357)—may serve as justifications for prejudice (Crandall et al., 2011), so that it is not surprising that stereotypes and prejudices correlate *empirically* (Dovidio et al., 1996). Since, *analytically*, stereotypes and prejudices are orthogonal concepts, both positive and negative prejudices can be built on more or less biased and, thus, incorrect stereotypes as well as on unbiased and, thus, correct stereotypes about any target group. Teachers' stereotypes about different groups of students I will examine in chapter 5.

That prejudice is not discrimination I put forth as a premise in chapter 2: While prejudice is an attitude, discrimination is about behavior. Knowing about somebody's prejudice is not equivalent to knowing about their discriminatory behavior. How the concepts of prejudice, stereotypes, and discrimination are linked and associated has to be addressed by both theory (see chapter 3) and empirical studies (see the discussion in the beginning of the present chapter).

4.2 Previous Research

There is not a lot of quantitative empirical research on prejudices of teachers in Germany or more generally their attitudes towards towards different ethnic or social groups. The few studies that do exist are—with regard to the research question raised in the beginning of this chapter—limited or biased due to the following reasons: First and foremost, they are often based on geographically limited convenience samples of students (Glock & Karbach, 2015; Hachfeld et al., 2015; Hachfeld et al., 2011; Hachfeld et al., 2012). As an example, I quantify the bias in Hachfeld et al. (2011) in section 4.4.1, where I return to the issue of biased and otherwise restricted samples in studies of teachers' attitudes and beliefs. Secondly, most published studies are limited due to the fact that their findings or reported descriptive numbers—if such numbers are reported at all—cannot either be interpreted in a meaningful absolute way or allow to compare relative biases towards students from different ethnic or social groups. Thirdly, not all studies that claim to investigate teachers' prejudices do so in the sense of the conceptualization above in section 4.1 but define or operationalize prejudice in a different way (see, e.g., Hachfeld et al., 2012, who, according to my definitions, measure stereotypes and beliefs instead).

4.2.1 Explicit Attitudes of Teachers in Germany

As the first in a series of studies by Axinja Hachfeld and colleagues, Hachfeld et al. (2011) report the results of a study with $N = 340$ teacher candidates and educational science students in Berlin, who turn out to score low on explicit measures of prejudice ($M = 1.76,\ SD = .57$; 5-point-scale, higher scores mean more negative prejudice) but high on explicit measures of both multicultural (mc) and egalitarian (eg) beliefs (mc: $M = 4.91,\ SD = .78$; eg: $M = 4.95,\ SD = .87$; 6-point-scales, higher scores mean beliefs that are more multicultural and egalitarian) (Hachfeld et al., 2011, p. 992). However, both the items assessing prejudice and the items assessing multicultural and egalitarian beliefs are hard to interpret in an absolute way and do not allow to compare attitudes toward different ethnic groups or groups of immigrants. Furthermore, making inferences from a student sample about the population of teachers in Germany would certainly be a bold move. In fact, I show below in section 4.4.1 that this sample indeed provides downwardly biased estimates of teachers' level of negative prejudice towards foreigners.

From the perspective of this chapter, the other studies by Hachfeld and colleagues have similar limitations. Hachfeld et al. (2012), for example, study various explicit attitudes and beliefs using a sample of $N = 433$ trainee teachers (German: *Lehramtsstudierende* or *Lehramtsanwärter/innen*) with and without immigrant background. However, what Hachfeld et al. (2012) call prejudice, is what I and many others (e.g., Ashmore & Del Boca, 1981; Ehrlich, 1973; Hilton & von Hippel, 1996; D. J. Schneider, 2004) call stereotypes or beliefs (see chapter 5). The items Hachfeld et al. (2012) use to measure what they call prejudice really investigate the respondents' beliefs about the interest, attention, thirst for knowledge, effort, and knowledge of students with immigrant background (Hachfeld et al., 2012, table 2). These statements can be more or less correct because they contain or make empirical statements about the reality of students with immigrant background. The results are a good case in point: Respondents with and those without immigrant background differ *the least* on these items while differing *the most* on items that are—among those implemented by Hachfeld et al. (2012)—probably closest to measuring prejudice, namely statements about whether or not respondents would enjoy teaching students with an immigrant background and students of different cultural background (Hachfeld et al., 2012, table 2). Taken together, Hachfeld et al. (2012)'s findings do not help to answer the questions raised in the beginning of the chapter.

There are other quantitative studies that implement explicit measures to measure teachers' prejudice that I have not discussed here since they, too, do not provide any

evidence with regard to the questions raised in the beginning of this chapter (e.g., Hachfeld et al., 2015; Wagner et al., 2001).

4.2.2 Implicit Attitudes of Teachers in Germany

Recently, a few studies have also investigated implicit attitudes of teachers in Germany. They, too, are often based on samples of students in education and not actual teachers. Glock and Karbach (2015), for example, report the results of an experimental study using three different implicit procedures (Affect misattribution procedure [AMP], Affective priming task [APT], Implicit Association Test [IAT]) on $N = 65$ German preservice teachers from two German universities. Respondents showed implicit preferences of ethnic majority students over ethnic minority (Turkish) students in all three measures (effect sizes: AMP: $d = .55$, APT: $d = .91$, IAT: $d = .93$). Differences were due to both negative implicit attitudes towards ethnic minority students (outgroup derogation) and positive attitudes towards ethnic majority students (ingroup favoritism). While the student sample is an obvious limitation of the study, it provides some evidence for implicit negative prejudice towards Turkish students of German teachers in general.

In a study with $n = 82$ elementary and $n = 82$ secondary school teachers that makes use of the Implicit Association Test (IAT), Glock and Klapproth (2017) investigate implicit prejudice towards ethnic minority students—that is, again, students of Turkish origin—of different gender. The results relevant for the present study are as follows: First, according to their reactions towards Turkish first names that were used as stimuli, both elementary and secondary school teachers show negative implicit prejudices to the disadvantage of Turkish students. IAT-effect sizes range from $D = .31$ to $D = 1.12$[15]. Secondly, the interaction effect of school type and students' gender is statistically significant. Its closer inspection reveals that while elementary school teachers show an implicit bias to the disadvantage of male compared to female ethnic minority students, secondary school teachers show the opposite pattern—that is a bias to the disadvantage of female compared to male students of Turkish origin. This result is mainly driven by the fact that elementary school teachers are more biased against male students of Turkish origin than secondary school teachers (difference: $d = .82,\ p < .001$). The implicit attitudes between both groups of teachers differ less

15 D is computed by dividing "the difference between test block means by the standard deviation of all the latencies in the two test blocks" (Greenwald et al., 2003, p. 201). Hence, its definition and interpretation is virtually the same as Cohen (1977)'s d that is usually calculated using the pooled standard deviation.

for female students (difference: $d = .39,\ p = .08$). Thirdly, the main effect of students' gender is far from any conventional significance levels ($p = .74$); thus, teachers are not prejudiced to the disadvantage of either girls or boys.

That the study is based on a sample of 164 teachers, not students, is an important advantage over other studies (Glock & Karbach, 2015; Hachfeld et al., 2015; Hachfeld et al., 2011; Hachfeld et al., 2012). However, that the sample is a convenience sample for which teachers were "recruited by undergraduates of the university"[16] by contacting "schools they were familiar with" (Glock & Klapproth, 2017, p. 80), is a limitation that should be kept in mind when interpreting the results. Another even more important limitation of the study by Glock and Klapproth (2017) is that the German names used in the IAT (male: Lukas, Finn, Niklas, Jonas, Tim, Paul; female: Leonie, Hannah, Julia, Emma, Marie, Sophie) probably carry a social class connotation that affects the results in an undesirable way. The names used may not simply signal an ethnic German background but a German upper middle class background instead, since the names are probably not an unbiased representation of the typical German student—students from lower or working class families tend to have other names. The names chosen may not only confound ethnic background with social class background but even exacerbate the socioeconomic differences that exist between the typical German student and the typical Turkish student. I will return to this point in chapter 6 in which I describe an experiment I conducted that also uses names as stimuli and that was designed to explicitly address this problem. Despite the limitations of their study, Glock and Klapproth (2017) provide additional evidence in favor of implicit biases of teachers in German schools against immigrants of Turkish origin.

4.3 Data

Especially through an ever increasing number of international and national studies in education, there are plenty of data sets available to the scientific community that contain variables on teachers in Germany. However, studies such as PISA, TIMSS, or PIRLS/IGLU do not contain measures of prejudice or attitudes towards immigrants in general or different ethnic group in particular. Also, when assessing prejudice in a study on education, teachers are probably aware that they are surveyed as teachers, which should make their professional role and identity salient to them, which, in turn, arguably increases the likelihood for answers that are socially desirable in the context

16 Probably from Wuppertal University, but this is not specified. The authors work at different institutions.

of education. This way, explicit measures of prejudice might underestimate teachers' true level of negative prejudice towards different ethnic or social groups. General social surveys, on the other hand, should not trigger the same mechanisms and, thus, should provide measures of prejudice that are less biased. Certainly, the sample size needs to be large enough to contain a sufficient number of teachers.

Therefore, I turned to data from general social surveys covering Germany, such as the German General Social Survey (*Allgemeine Bevölkerungsumfrage der Sozialwissenschaften*, ALLBUS), the European Social Survey (ESS), the European Values Study (EVS), the International Social Survey Programme (ISSP), and the World Values Study (WVS). Except the ALLBUS, none of the studies assesses attitudes towards particular ethnic groups but instead measures negative prejudice, stereotypes, and other attitudes towards racial or ethnic minorities more generally. Such questions mostly aim at "foreigners", "foreign workers" (EVS/WVS), or "immigrants" (EVS/WVS), and sometimes more marginalized groups such as "gypsies" (EVS/WVS). This is why I make use of the ALLBUS for the analyses in this chapter.

4.3.1 The ALLBUS

In addition to more general questions on immigrants and foreigners similar to those asked in ESS, ISSP, and EVS/WVS, the ALLBUS asks several questions that assess the *social distance* of the respondent towards Italians, Eastern Europeans of German descent, asylum seekers, Turks, and Jews. The corresponding questions were asked in 1996, 2006, and 2016. Since 2006 is closer to the data collection for the experiment discussed in chapter 6, I will use the data from ALLBUS 2006 for the analyses in this chapter.

4.3.2 Social Distance: A Global Measure of Prejudice

Social distance measures were among the first and have been among the most popular explicit global measures of prejudice (see Correll et al., 2010; Ehrlich, 1973, for overviews). The concept of social distance was introduced to the study of racial attitudes and race relations by R. E. Park (1924), who defined it as "the grades and degrees of understanding and intimacy which characterize personal and social relations generally" (R. E. Park, 1924, p. 339). Prejudice, R. E. Park (1924, p. 339) suggested, was the "more or less instinctive and spontaneous disposition to maintain social distances".

Bogardus (1925) operationalized the social distance concept and empirically tested it using a mix of quantitative and qualitative methods. However, based on pretest

results for 60 items, Bogardus (1933) suggested a core of seven equidistant social situations, each represented by one item, along which social distance towards various ethnic or social groups is reported by respondents. The social situations vary according to their degree of intimacy and range from "(1) would marry" over "(4) would have several families in my neighborhood" to "(7) would have live outside my country" (Bogardus, 1933). Since then the measure has been used—in its original form or adapted forms—in many studies in different countries (e.g., Bobo & Hutchings, 1996; Bogardus, 1958; Böltken, 2000; Ganter, 2003; Hill, 1984; Kleinert, 2004; Parrillo & Donoghue, 2005, 2013; Smith & Dempsey, 1983; Stangor et al., 1991; Steinbach, 2004; Storm et al., 2017; H. C. Triandis & Triandis, 1960, 1962).

Social distance in the ALLBUS 2006

In the ALLBUS 2006, social distance is measured towards the following groups: Italians, Eastern European immigrants of German descent ("deutschstämmiger Aussiedler aus Osteuropa"), Asylum seekers, Turks, and Jews[17]. For each group, two of the seven core items from Bogardus (1933) are administered:

- How pleasant or unpleasant would it feel to you to have members of the following groups as neighbors? How pleasant or unpleasant would it feel to have …
  - an Italian as neighbor?
  - an Eastern European immigrant of German descent as neighbor?
  - …
- And what if a member of one of these groups were to marry into your family? To what extent would this feel pleasant or unpleasant to you? To what extent would it feel pleasant or unpleasant to you, …
  - if an Italian were to marry into your family?
  - if an Eastern European immigrant of German descent were to marry into your family?
  - …

The response scale ranges from -3 ("very unpleasant") over the unlabeled midpoint of 0 to $+3$ ("very pleasant") for both questions.

Of the various groups I will investigate social distance of teachers in Germany towards Turks, because they are the ethnic group I mainly focus on in this study of discrimination in German education. Mainly to have meaningful standards of comparison, I will also investigate teachers' social distance towards Italians and Eastern Eu-

17 In order of appearance in questionnaire.

ropean immigrants of German descent, henceforth simply "Eastern Europeans'. Both asylum seekers and Jews I will leave aside. The group of asylum seekers is a very small, heterogeneous, and a quickly changing group so that I think not much is learned by including them here. Jews in Germany are also a small group and social distance towards them tap on different dimensions than the three ethnic groups I have selected for comparison.

4.4 Analytic Strategy[18]

4.4.1 Identifying Teachers in Data from General Social Surveys

In data from general social surveys, such as the ALLBUS, teachers—just like any other occupational group—can be identified using the International Standard Classification of Occupations (ISCO). While the most recent classification scheme is ISCO-08, the ALLBUS 2006 features the version that was up-to-date at the time, ISCO-88 (International Labour Organization, 1990). ISCO-88 distinguishes ten *major groups,* two of which contain teachers and other educators. Teachers or educators that hold tertiary degrees are classified as "professionals" in major group 2 that "includes occupations whose main tasks require a high level of professional knowledge and experience in the fields of physical and life sciences, or social sciences and humanities" (International Labour Organization, 1990, section "Summary of Major Groups"). This applies to all regular teachers in German elementary and secondary schools and, of course, lecturers, readers, and professors at universities. These "teaching professionals" are classified as *sub-major group* 23.

Teachers or educators that do not hold tertiary degrees fall into major group 3 that comprises occupations "whose main tasks require technical knowledge and experience" (International Labour Organization, 1990, section "Summary of Major Groups") in the same fields as above. In Germany, teachers and educators without a tertiary degree include educators in preschool and kindergarten—the main institutional education and care settings for children below the age of 6. These "teaching associate professionals" are classified as sub-major group 33. Education related occupations are also classified in sub-major groups 12, 13, and, 51.

The next level below the sub-major group level in ISCO-88 is the *minor group*. The last and, thus, finest level is the *unit* on which 390 occupational groups are uniquely identified by a 4-digit code. Typically, these groups consist of more than one occu-

18 Syntax to replicate all empirical analyses in this chapter is available at https://osf.io/dqtkg/.

pation (International Labour Organization, 1990, section "Design and Structure"). I describe the operationalization of the *teacher* variable in the next section.

Operational definitions

This study focuses on discrimination in elementary education. Therefore, investigating the attitudes of elementary school teachers would be a priority over more general operationalizations of what a *teacher* is. However, in the ALLBUS 2006 there are only $N = 7$ teachers that can clearly be identified as elementary school teachers (see appendix B). So, to reliably learn about teachers' attitudes towards different ethnic groups, I operationalize teachers in a broader way and look at the attitudes of *school teachers*. This operationalization results in a variable that equals 1 for all respondents generically classified as teachers holding a tertiary degree as well as all respondents classified particularly as secondary and elementary school teachers as well as those working in special education and 0 for everybody else. The table in B shows the operationalization and the number of observations per ISCO-88 unit. In sum, I can identify $N = 51$ respondents as school teachers in the ALLBUS 2006. School teachers are between 25 and 65 years old ($M = 47.7,\ SD = 9.74$), a majority of 68.7% is female and 12.5% is from East Germany[19].

I assess the sensitivity of my results by also looking at the attitudes of an even larger group of respondents: I construct a variable that—in addition to school teachers—also equals 1 for those teaching at colleges or universities, those that work as teachers but only hold a secondary degree, as well as those who work in preschool and kindergarten and hold either a tertiary or secondary educational degree. In fact, of the 25 respondents I gain in this second operationalization 20 work in pre-primary education (ISCO-88 units 2332 and 3320, see appendix B). Therefore, the group of *all educators* ($N = 76$) comprises those respondents who personally teach, educate, or take care of children or students of different ages in institutional settings. Respondents in this group are between 23 and 65 years old ($M = 46.1,\ SD = 10.69$), a majority of 74.1% is female and 20.4% is from East Germany[20].

While both operationalizations have less observations than we might wish for, note that quantitative studies in education that focus on variables on the level of teachers often feature similar or even lower numbers of teachers. Recent examples of

19 Means, standard deviations, and proportions are calculated using weights to account for oversampling of respondents living in East Germany.

20 As for school teachers, means, standard deviations, and proportions are calculated using weights.

small-N studies on teachers come from different lines of research and include studies on teachers' implicit attitudes (Bergh et al., 2010; Glock & Karbach, 2015; Glock & Klapproth, 2017), studies that investigate teachers' expectations as self-fulfilling prophecies (Lorenz et al., 2016), research on teachers' diagnostic competence (Artelt & Rausch, 2014; Karing et al., 2011), and, last but not least, studies that experimentally and, thus, more directly investigate discrimination in education (Sprietsma, 2013). Note that most of these studies not only investigate simple summary statistics but often conduct more or less complex multivariate analyses on these samples, meaning that the number of observations per cell and, therefore, statistical power is reduced further.

A key reason for the relatively small number of observations in these studies is that it is rather difficult and, thus, costly to draw probability samples of teachers. A possible solution is to investigate beliefs or attitudes of pre-service or beginning teachers that are still enrolled as students or educational science students more broadly and to use regional convenience samples (e.g., Hachfeld et al., 2015; Hachfeld et al., 2011; Hachfeld et al., 2012). This way, samples of $N > 100$ observations can be achieved more easily. However, while such samples might be considered appropriate for testing the reliability and validity of new instruments, they are a severe limitation if interest lies in quantifying the number of prejudiced teachers. This is, because these samples are biased with regard to variables that are known to determine the valence of attitudes. These variables include age, gender, and region of residence. Note that it is all the more astonishing that some of the small-N studies cited above are also conducted using samples that are restricted in these ways—for example, featuring students only (e.g., Glock & Karbach, 2015).

The bias of convenience samples

To assess the bias of geographically limited convenience samples of students compared to probability samples of teachers I exploit the fact that Hachfeld et al. (2011, study 2) implement four items from an item-battery repeatedly used in the ALLBUS to assess prejudice against "foreigners". The original German item wordings and a translation into English can be found in appendix A. The sample of Hachfeld et al. (2011) comprises $N = 340$ students (233 women) from a German university of which "79% ($n = 254$) were of German nationality and 21% ($n = 68$) had an immigrant background" (Hachfeld et al., 2011, p. 992)[21]. Participants were 19 to 55 years old

21 I quote Hachfeld et al. (2011), since the two categories—being of German nationality and having an immigrant background—are, of course, not mutually exclusive.

($M = 25,\ SD = 5$) and were either teacher candidates ($n = 266,\ 81\%$) or students of educational science ($n = 62,\ 19\%$).

Because Hachfeld et al. (2011) changed the original 7-point scale into a 5-point scale, I cannot directly compare the mean ($M = 1.76$) and standard deviation ($SD = .57$) reported by Hachfeld et al. (2011) to the mean and standard deviation of school teachers in the ALLBUS 2006 ($M = 2.82,\ SD = 1.10$; weights apply). Higher numbers stand for more negatively prejudiced attitudes towards foreigners. To compare the mean responses of participants in Hachfeld et al. (2011) and the school teachers identified in the ALLBUS 2006, I calculate Cohen (1977)'s d by taking the difference of mean to scale midpoint divided by the sample standard deviation for the respective group[22]. This calculation yields an effect size of $d = (3 - 1.76)/.57 = 2.18$ for the numbers reported in Hachfeld et al. (2011). That is, students that participated in Hachfeld et al. (2011) lie over 2 standard deviations away from the scale midpoint towards the less prejudiced pole of the scale. For the school teachers identified in the ALLBUS 2006, the corresponding calculation yields $d = (4 - 2.82)/1.10 = 1.07$, which is about half the size into the same direction from the scale midpoint.

In summary, the results of the calculation above suggest that, as expected, geographically limited convenience samples of students are selective samples that should not be trusted as an unbiased source of evidence on the distribution and valence of teachers' attitudes or beliefs.

4.4.2 Absolute and Relative Measures of Prejudice

I construct and calculate both absolute and relative measures of prejudice. Absolute measures of social distance are meaningful since they provide an estimate for the valence and distribution of negative prejudices towards different ethnic or social groups. However, relative measures are even more important because, strictly speaking, only the relative distances between different groups allow us to hypothesize about discrimination understood as causal effect of, say, ethnicity, in the counterfactual sense. Therefore, it could be argued—and I have (see, e.g., chapter 3)—that absolute measures of prejudice, too, are only relevant for the study of discrimination relative to measures of prejudice towards other groups or an assumed baseline of negative or positive prejudice towards the ethnic majority or teachers' in-group.

Recall (or see section 3.1.1 again) the discussion of Becker (1957/1971)'s taste for

22 Cohen (1977, p. 20) defined the effect size d as the mean difference of two groups or populations divided by the standard deviation of either group or population that are assumed equal.

discrimination, d, that, if $d_i > 0$ toward a particular group, increases the net costs of employing a worker from that group for employer i. However, to understand what a positive taste for discrimination against a particular group means for an employer's behavior—or teacher's behavior, for that matter—towards members of this group, we need to be explicit about the reference group—that is, the counterfactual causal state—and the corresponding d_i for that group. Note further that similar arguments can be derived from social identity theory and other approaches resting on both ingroup favoritism and outgroup derogation that may vary in intensity between different in- and out-groups. The importance of relative distances between groups including the ethnic majority or other ingroups is also acknowledged in implicit measures such as the IAT, where the valence of prejudice against outgroups is operationalized as difference (in differences) to a reference group.

Absolute measures

As a meaningful absolute measure of teachers' negative prejudice towards the different ethnic groups—that is, Turks, Eastern Europeans, and Italians—I calculate two indicators: First, I calculate the proportion of respondents that finds it *either* unpleasant if a member of the ethnic group in question were to marry into their family *or* unpleasant having members of the respective group as neighbors. This is achieved by creating a dummy variable that is 0 for those respondents who report to be either indifferent (0 on the response scale) or to feel pleasant (+1/+2/+3 on the response scale) towards *both* social situations, and 1 for everybody else.

As a more conservative absolute estimate of negative prejudice, I calculate, secondly, the proportion of respondents that reports to find *both* situations unpleasant: a member of the respective group marries into the respondents family *and* members of the respective group are neighbors of the respondent. The corresponding dummy variable is 0 for all respondents who report to be either indifferent (0 on the response scale) or to feel pleasant (+1/+2/+3 on the response scale) towards *either* social situation, and 1 for everybody else. Clearly, respondents that score 1 on this measure have stronger negative prejudice against the respective group than the respondents that score 1 on the measure above. Thus, the proportions of teachers and educators holding negative prejudice should be lower according to this measure. Also, since the scale by Bogardus (1933) was set up as a Guttman scale, it can be expected that those who feel unpleasant about having a member of a particular group as neighbor, would most likely also feel unpleasant if a member of the same group were to marry into

their family. Since my interest is more of substantive than of technical nature, I will not investigate this further.

From both dummy variables, I calculate the proportions of three groups of respondents—school teachers, educators, and all respondents—reporting negative prejudice against the different ethnic groups using weights that correct for the oversampling of East Germans.

Relative measures

I propose three relative measures of negative prejudice: First, I will simply take the *differences in proportions* of negative prejudice towards different groups as operationalized above. I will test whether these differences are statistically significant. Weights will be applied to address the oversampling of respondents living in East Germany.

For the two remaining measures, I construct a sum score of prejudice towards each ethnic group by adding the responses to the two social distance questions asked about each group. This yields a prejudice sum score with corresponding mean (M), standard deviation (SD), and Cronbach's alpha (α) per ethnic group—that is, Turks, Eastern Europeans, Italians—and respondent group. Based on these sum scores, secondly, I calculate *t-tests* to obtain mean differences with p-values and confidence intervals. Here, too, I apply weights.

Thirdly, also based on the sum scores, I calculate *effect sizes* for mean differences in teachers' prejudice towards different groups to enable comparisons "free of our original measurement unit" Cohen (1977, p. 20). Cohen (1977) suggested the effect size d, defined as the mean difference of two groups or populations divided by the standard deviation of either group or population that are assumed equal. I estimate d following Cohen (1977, pp. 66–67) as

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s^*}, \tag{4.1}$$

where

$$s^* = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{4.2}$$

and where s_1^2 and s_2^2 are the unbiased sample variances of x_1 and x_2. Note that weights are applied throughout and, thus, n_1 and n_2 are the weighted number of observations[23].

23 The weighted number of observations deviates only slightly from the actual number of ob-

4.5 Results

4.5.1 Proportion of Teachers with Negative Prejudice

Table 4.1 shows the proportions of school teachers, all educators, and all respondents holding negative prejudices against Turks, Eastern Europeans, and Italians along with the number of observations (n) and standard errors (SE) for two different operationalizations of negative prejudice. While the absolute numbers speak for themselves, their stability over the two different groups of teachers—*school teachers* versus *all educators*—lends credibility to the absolute numbers themselves as well as to the following analyses regarding the differences and relative distances between groups. As is apparent from the numbers in table 4.1, the ranking of the groups is stable over all operationalizations of prejudice and all groups of respondents: Most negative is the prejudice towards Turks, followed by Eastern Europeans, and then Italians, against whom the two groups of teachers report virtually no negative prejudice no matter the operationalization (0%-3%). Compared with—depending on the operationalization—25 to 44 percent of teachers reporting negative prejudice against Turks and 12 to 28 percent reporting negative prejudice toward Eastern Europeans, this leads to a clear "ethnic hierarchy" (Hagendoorn, 1995) in social distances towards the three groups.

All contrasts between different ethnic groups are statistically significant with $p < .05$ for all groups of respondents and operationalizations of prejudice, except the difference between Turks and Eastern Europeans in the more conservative measure of prejudice ("AND" in table 4.1) for the smaller group of school teachers; this contrast is only significant on the 10% level ($t = 1.68,\ p = .099$).

4.5.2 Mean Differences and Effect Sizes

Virtually the same results are obtained by turning to mean differences between the sum scores of prejudice towards the different ethnic groups. For school teachers, sum scores yield the following numbers: Prejudice towards Turks ($M = -.05,\ SD = 2.74,\ \alpha = .82$), prejudice towards Eastern Europeans ($M = .56,\ SD = 2.46,\ \alpha = .73$), prejudice towards Italians ($M = 3.03,\ SD = 2.35,\ \alpha = .73$). For all educators the numbers are very similar: Prejudice towards Turks ($M = -.07,\ SD = 2.95,\ \alpha = .82$), prejudice towards Eastern Europeans ($M = .97,\ SD = 2.65,\ \alpha = .73$), prejudice towards Italians ($M = 2.92,\ SD = 2.37,\ \alpha = .73$). Again, all possible contrasts between

servations (e.g., $n_{weighted} = 49.13$ instead of $n_{real} = 48$ for the sum score of school teachers' prejudice against Turks).

Table 4.1: Proportion of school teachers, all educators, and all respondents holding negative prejudices against different ethnic groups.

			AND		OR	
Respondents	Target group	n	Proportion	(SE)	Proportion	(SE)
School teachers	Turks	49	.25	(.06)	.44	(.07)
	Eastern Europeans	49	.15	(.05)	.28	(.06)
	Italians	49	.00		.02	(.004)
All educators	Turks	70	.26	(.05)	.44	(.06)
	Eastern Europeans	72	.12	(.04)	.24	(.05)
	Italians	72	.01	(.01)	.03	(.02)
All respondents	Turks	3316	.34	(.008)	.51	(.009)
	Eastern Europeans	3310	.20	(.007)	.34	(.008)
	Italians	3321	.03	(.003)	.10	(.005)

Source: ALLBUS 2006, own calculations applying weights.

AND: marriage *and* neighbor unpleasant; OR: marriage *or* neighbor unpleasant

ethnic groups are statistically significant ($p < .01$) for both school teachers and all educators except the contrast between Turks and Eastern Europeans for the smaller group of school teachers that only reaches significance on the 10% level ($t = -1.73,\ p = .09$).

Even more so than differences in proportions, effect sizes help to substantiate interpretations of how large the size of a group difference really is. Again, I prefer to let the numbers speak for themselves—instead of applying arbitrary schemes that supposedly help to decide when a difference is small, medium, or large (Cohen, 1977, pp. 24–27)[24]. Effect sizes for all contrasts between ethnic groups and both groups of teachers (d_1: school teachers; d_2: all educators) are: Turks-Eastern Europeans ($d_1 = -.23,\ d_2 = -.37$), Turks-Italians ($d_1 = -1.21,\ d_2 = -1.12$), Eastern Europeans-Italians ($d_1 = -1.03,\ d_2 = -.77$). As with all other measures, the relative ranks of the groups according to effect size differences are such that Turks are target of the most negative prejudice, followed by Eastern Europeans, followed, in turn, by Italians.

24 Should the reader be interested or in need of a reminder or both, Cohen (1977, pp. 24–27) suggested to call $d = .2$ small, $d = .5$ medium, and $d = .8$ large, even though he acknowledged that "a certain risk inherent in offering conventional operational definitions for these terms for use in power analysis in as diverse a field of inquiry as behavioral science" (Cohen, 1977, p. 25).

4.6 Summary and Conclusion

In this chapter, I have been concerned with teachers' prejudices. Understood as an attitude towards and, hence, as an evaluation of a particular group or category of people, prejudice is both theoretically and empirically a key and, in fact, the most important determinant of discriminatory behavior.

Previous research on teachers' prejudice has not been particularly enlightening because of several shortcomings of which the most important are that these studies typically rely on geographically limited convenience samples of students, that their findings or reported numbers can neither be interpreted in a meaningful absolute way nor used to calculate differences between various ethnic or social groups, and sometimes use different terminology than the one proposed in this chapter and investigate stereotypes or other forms of beliefs instead of prejudice.

In this chapter, I have used data from the German general social survey, ALLBUS, collected in 2006, to show how biased the samples of previous studies are. Judging by effect sizes, studies based on geographically limited convenience samples of students appear to underestimate teachers' prejudice towards immigrants by about one standard deviation. Using the same data set, I address this and the other limitations of previous studies by investigating teachers' explicit attitudes towards different ethnic groups living in Germany. Using the 4-digit ISCO-88 unit code, I am able to identify a group of $N = 51$ teachers in elementary or primary and secondary schools and a group of $N = 76$ educators that additionally include lecturers and professors at universities as well as educators in preschool and kindergarten. Obviously, analyses based on the larger group have higher statistical power and allow to assess the sensitivity of my results for school teachers—the group I am primarily interested in.

To examine teachers' prejudice, I make use of measures of social distance in two social situations and towards three different ethnic groups, namely Turks, Eastern Europeans of German descent, and Italians. I construct and calculate both absolute and relative measures of prejudice. As for an absolute measure, I calculate the proportions of teachers with negative prejudice towards the three ethnic groups. Using two different operationalizations—one more, one less conservative—I can show that negative prejudice against Turks is pretty widespread among teachers in German schools but less so against Eastern Europeans and virtually non-existent against Italians. More important than absolute numbers are the relative distances between groups. Based on sum scores of social distance measures, I find differences in the magnitudes of bias against the groups that result in effect sizes of $d > 1$ for the contrast between Turks and Italians, and around $d = .8$ and $d = 1$ for the contrast between Eastern Europeans and Italians.

In conclusion, my analyses of explicit prejudice confirm the findings of Glock and Klapproth (2017) for implicit prejudice, namely that teachers in Germany hold attitudes that are biased against Turks. What it adds is that teachers also hold negative prejudice against Eastern Europeans, but not Italians. To put the findings into broader perspective, note that all my analyses provide evidence for a very clear ethnic hierarchy that—from top to bottom—looks as follows: Italians > Eastern Europeans > Turks. This pattern confirms findings in other studies for Germany (e.g., Ganter, 2003; Kleinert, 2004; Steinbach, 2004) but also other countries; especially the finding of Turks at or near the bottom of the hierarchy has been replicated for different western countries at different historical times (Bogardus, 1925; Hagendoorn, 1995; Hraba et al., 1989).

With regard to discrimination in German education on both individual and group level and, eventually, the explanation of inequality between students of different ethnic origin, my findings suggest that prejudice-based or taste-based discrimination to the disadvantage of students with Turkish as well as Eastern European background cannot be ruled out. In contrast, the disadvantage of Italian students (e.g., Kristen, 2002; Kristen & Granato, 2007; Olczyk, 2016) is probably not due to such a form of discrimination.

There are some noteworthy limitations of the study presented in this chapter. First, even though the sample sizes for the two groups of teachers I distinguished are of comparable size as the samples of other studies on teachers' attitudes or beliefs (e.g., Glock & Karbach, 2015; Lorenz et al., 2016), a replication using a larger sample of teachers would be desirable. Secondly, using data from a general social survey, we cannot know whether teachers actually teach students of Turkish, Eastern European, or Italian origin. The experiment in chapter 6 will address this problem. Thirdly, a point I briefly mentioned when discussing the design of Glock and Klapproth (2017)'s study is that measures of ethnic prejudice may confound ethnic and social class prejudice (e.g., Blalock, 1967, pp. 199–203), since the ethnic groups towards which prejudice is expressed usually vary regarding their endowment with cultural, economic, and social resources—in particular compared to the ethnic majority. For the analyses and findings in this chapter, though, I suggest that social class differences should not play an important role in explaining the differences in teachers' social distances towards ethnic groups since socioeconomic differences between ethnic groups in Germany are simply not pronounced enough (Büchel & Frick, 2004; Kalter, 2008; Kogan, 2007). Fourthly, just like previous research, I could not examine teachers' prejudice against different social classes or against men and women or boys and girls, for that matter.

5 Stereotypes of German Teachers[25]

> Stereotypes systematically affect how people perceive, process information about and respond to, group members.
>
> *(Dovidio et al., 2010)*

In this chapter, I am concerned with stereotypes of German teachers towards different groups of students. In line with many contributions to the social cognition literature—the dominant field within social psychology since several decades—I define a stereotype as *a belief or a set of beliefs about the characteristics, attributes, or behaviors of a particular group or category of people* (see, e.g., Hilton & von Hippel, 1996, p. 240; Ashmore & Del Boca, 1981, p. 16; D. J. Schneider, 2004, p. 24; Ehrlich, 1973, p. 20).

It is vital for any study of discrimination in education to understand the mechanics, contents, and valence of teachers' stereotypes, since stereotypes are the key determinant of discriminatory behavior in some of the most import theories of discrimination (e.g., Aigner & Cain, 1977; Fiske et al., 1999; Gaertner & Dovidio, 1986; Phelps, 1972) and have been shown to serve particular functions in perceiving, storing, and retrieving information in numerous studies. It has been shown that stereotypes and their use in interpersonal interactions are connected to a largely inevitable and automatic process of categorizing people on the basis of biological and social cues (Allport, 1954; Fiske, 1998; Fiske et al., 1999). Also, people tend to seek (Darley & Gross, 1983; Snyder & Swann, 1978), encode (Bodenhausen & Lichtenstein, 1987), recall (Bodenhausen & Lichtenstein, 1987), and interpret (Darley & Gross, 1983) information in a stereotype-consistent way. Eventually, stereotypes may influence the way people judge and treat other people and, therefore, lead to discrimination that disadvantages individuals or, potentially, whole groups. Understood as hypotheses about outgroups or ingroups (D. J. Schneider, 2004, pp. 197–228; Snyder & Swann, 1978), stereotypes may turn into self-fulfilling prophecies (Jussim, 1989; Jussim et al., 1996; Jussim & Harber, 2005; Jussim, Robustelli, et al., 2009; Lorenz et al., 2016) or affect the behavior of targets of stereotypes such that members of the targeted group behave in a way that tends to confirm the stereotype and, thus, affect real-world outcomes in, for example, education (e.g., Lee et al., 1995; Nguyen & Ryan, 2008; Schmader et al., 2008; Stricker et

25 This chapter is based on joint work with Melanie Olczyk and Georg Lorenz (Wenz et al., 2016).

al., 2015). Therefore, the first research question I am going to address in this chapter is simply this: Can we *validly measure* teachers' stereotypes about different groups of students?

However, the links from categorization to stereotype activation, from stereotype activation to application, and from stereotype application to discrimination are not automatic and inevitable processes that occur regardless of other, moderating, factors (see, e.g. Macrae & Bodenhausen, 2000). Therefore, we need theories of discrimination (see chapter 3) that tell us under which conditions stereotypes do turn into discriminatory behavior and under which conditions they don't and under which conditions discriminatory behavior based on stereotypes disadvantages whole groups (see sections 2.4.1 and 3.1.2 for the distinction between individual discrimination and group discrimination). Following different theoretical approaches (e.g., Aigner & Cain, 1977; Fiske et al., 1999; Gaertner & Dovidio, 1986), stereotypes—i.e., their content or valence—need to differ between groups, in order to explain individual discrimination. Therefore, the second question I seek to answer in this chapter is whether teachers hold stereotypes about different groups of students that *differ* between groups? Although there are other conditions under which group discrimination arises from individual discrimination, with regard to the nature of stereotypes, it is the erroneous perception of group differences that leads to group discrimination (Aigner & Cain, 1977; England & Lewin, 1989). Therefore, the third question I intend to address is whether teachers' stereotypes are *accurate or biased* and, if they are biased, to which groups' (dis)advantage?

5.1 Conceptualizing Stereotypes

Research on stereotypes has a long history: From Lippmann (1922)'s "pictures in our heads" metaphor until today's multifaceted perspectives on the term, definitions of stereotypes abound. I will—unlike in chapter 2—not provide an in-depth discussion of many different conceptualizations. But since this chapter deals with the development of a new measure of teachers' stereotypes, it is necessary to provide a more detailed account of how stereotypes have been conceptualized than for prejudice in chapter 4. Therefore, I will—as concise as possible—explain why I chose a particular definition over others, that is, why I find it more useful in an empirical study on discrimination in education than alternative conceptualizations.

For more detailed reviews of different conceptualizations of stereotypes see e.g., D. J. Schneider (2004, pp. 14–26), Ashmore and Del Boca (1981), Leyens et al. (1994, pp. 9–18), Nelson (2006, pp. 4–7).

5.1.1 Useful and Not so Useful Definitions

Recall that I—in line with key contributions to the literature (e.g., Hilton & von Hippel, 1996, p. 240; Ashmore & Del Boca, 1981, p. 16; D. J. Schneider, 2004, p. 24; Ehrlich, 1973, p. 20)—define a stereotype as *a belief or a set of beliefs about the characteristics, attributes, or behaviors of a particular group or category of people.* Defined in this way, a stereotype is a cognitive structure that links knowledge to a category of people (Bless et al., 2004, p. 53; Macrae & Bodenhausen, 2000).

My definition is very close to one of the most widely used definitions, namely the one proposed by Ashmore and Del Boca (1981):

> Thus, we propose the following as the core meaning of the term "stereotype": *A set of beliefs about the personal attributes of a group of people.* (Ashmore & Del Boca, 1981, p. 16, their emphasis)

Also very similar is the definition by D. J. Schneider (2004), who also provides a key argument in favor of his—and my—definition:

> The most basic definition I can offer, the one with the fewest constraining assumptions, is that *stereotypes are qualities perceived to be associated with particular groups or categories of people*. (D. J. Schneider, 2004, p. 24, his emphasis)

D. J. Schneider (2004) argument that should be a key argument for any researcher to pick a definition, is that it is a basic or general one. It is a definition that does not constrain stereotypes in unnecessary ways to be a particular subset of beliefs about groups or empirically work in particular ways. I discuss these subsets and empirical mechanisms in the paragraphs below.

True or false?

One of the oldest debates around the term has been concerned with the question of whether stereotypes should be conceptualized as incorrect per se. Katz and Braly (1933, 1935), who conducted the first systematic empirical studies on stereotypes, defined a stereotype as "a fixed impression, which conforms very little to the facts it pretends to represent" (Katz & Braly, 1935, p. 181). Similarly, Allport (1954) suggests that a stereotype is "an exaggerated belief associated with a category" (p. 191) and, hence, rules out by definition that a stereotype can be "a valid generalization" (p. 192). In contrast, some 20 years later, Ehrlich (1973) was much less restrictive and allowed stereotypes to also be correct, in referring to stereotypes as "a set of beliefs and disbeliefs about any group of people".

Over time it has become the "standard viewpoint" (Hilton & von Hippel, 1996, p. 240) to allow stereotypes to contain more or less accurate beliefs—that is, empirically, they can be exactly right or completely wrong or anything in between. My conceptualization, that was also adopted for the item development in the NEPS (Wenz et al., 2016), is consistent with this standard viewpoint; only from this standard viewpoint it makes sense to ask and empirically investigate the question whether and, if so, how biased teachers' stereotypes are.

Individual or cultural?

Furthermore, different forms of stereotypes have been discussed. One important distinction separates personal or individual from cultural stereotypes (Ashmore & Del Boca, 1979, 1981; Gardner, 1973). Ashmore and Del Boca (1981), for instance, suggested "that the term 'stereotype' should be reserved for the set of beliefs held by an individual regarding a social group and that the term 'cultural stereotype' should be used to describe shared or community-wide patterns of beliefs" (Ashmore & Del Boca, 1981, p. 19).

Especially earlier contributions conceptualized stereotypes as shared or consensual beliefs about the characteristics or attributes of groups. That is, cultural consensus of some form was a necessary condition for a belief to be called a stereotype. The studies of Katz and Braly (1933), Katz and Braly (1935), for example, implicitly built on this premise. Vinacke (1957) argued that these and other earlier studies on stereotypes were based on an understanding of stereotypes as "a collection of trait-names upon which a large percentage of people agree as appropriate for describing some class of individuals (Vinacke, 1957, p. 230). Even Gardner (1973), who introduced an individual perspective by focusing on the individual stereotyper and, thus, on the process of stereotyping, upheld the "traditional definition of stereotypes as consensual beliefs about the characteristics of ethnic groups" (Gardner, 1973, p. 134).

Around that time—i.e., in the 1970s (D. J. Schneider, 2004, p. 22)—more and more scholars changed to a more individual perspective on stereotypes and processes of stereotyping (Ashmore & Del Boca, 1979; Ehrlich, 1973, e.g.,). In recent years, it has been a shared belief among social psychologists and other social scientists that stereotypes should be conceptualized as beliefs held by individuals (see, e.g., Bordalo et al., 2016; Dovidio et al., 2010; Hilton & von Hippel, 1996; Nelson, 2006; D. J. Schneider, 2004). Note, however, that this perspective does not imply that there is no social or cultural aspect to stereotypes: Indeed, individual level and societal level variables and processes do interact in determining contents and valence of individual beliefs—i.e.,

stereotypes. Stereotypes held by one individual can but do not have to be shared by others and widely shared beliefs are likely to be known by those who do not endorse them explicitly (Devine, 1989, p. 5). Conversely, knowing about cultural stereotypes might be enough to build implicit associations that are different from explicit beliefs in that they are hard to control, automatic constructs (Devine, 1989).

While I am not particularly interested in stereotypes held by single individuals—i.e., single teachers—I follow the logic of methodological individualism and aim at measuring individual stereotypes. This position was also adopted in the NEPS and the question whether and how to aggregate the individual beliefs—e.g., in a statistical model or by defining a criterion for a cultural consensus among teachers—was left to the data user (Wenz et al., 2016, p. 5).

Explicit or Implicit?

In recent years the distinction between explicit beliefs and implicit associations has been the most important and most debated in research on stereotypes and attitudes (see, e.g., Cunningham et al., 2001; Fazio & Olson, 2003; Gawronski & Bodenhausen, 2006; Greenwald & Banaji, 1995; Greenwald et al., 1998). While some scholars suggest a distinction between implicit and explicit attitudes and stereotypes (e.g., Cunningham et al., 2001; Greenwald & Banaji, 1995; Greenwald et al., 1998), others distinguish between explicit and implicit *measures* of stereotypes, attitudes, and the like (Fazio & Olson, 2003, pp. 302–303). In line with Wenz et al. (2016), I take the latter position and will elaborate on this distinction below in section 5.2.

Stereotypes and related constructs

Before I continue with a discussion of how stereotypes have been measured, let me briefly summarize the differences between stereotypes and related constructs, namely stereotyping, prejudice, and discrimination. While the latter two have been discussed in chapters 2 and 4, I have hitherto not elaborated on the difference between stereotypes and stereotyping.

In line with the literature, I explicitly distinguish stereotypes from stereotyping and conceptualize *stereotyping* as the *process of applying a stereotype in any judgment, impression, or expectation formation—be it towards an individual or a number of individuals* (see, e.g., Hilton & von Hippel, 1996; Leyens et al., 1994; Macrae & Bodenhausen, 2000; Zarate & Smith, 1990). The distinction matters more than is obvious on first sight, because it implies that stereotypes themselves cannot be measured by assessing im-

pressions or judgments of individuals. For example, teachers' stereotypes about the educational abilities of different groups of students cannot be measured by asking teachers to rate the educational ability of single students and then assess potential bias in these ratings by virtue of particular group characteristics such as ethnicity, social class, or gender.

As laid out in chapter 4, by the term *prejudice* I refer to *an attitude toward a particular group or category of people* (see also, e.g., Correll et al., 2010; Ehrlich, 1973; D. J. Schneider, 2004). Since by an attitude, I mean "general evaluations of people, objects, and issues" (Fazio & Petty, 2008, p. 1), prejudice is an evaluation of a group or category of people. In contrast, stereotypes, understood as beliefs, lack any evaluative component. Thus, the key difference is that, while stereotypes can be correct or—probably more often—more or less incorrect, prejudices can't. An evaluation is positive or negative but neither false or inaccurate nor true or accurate, respectively. However, stereotypes—sometimes referred to as the cognitive component of prejudice (Dovidio et al., 2010, p. 5; Fiske, 1998, p. 357)—may serve as justifications for prejudice (Crandall et al., 2011). Therefore, a negative prejudice can certainly be built on a biased and, thus, incorrect stereotype.

Discrimination, as defined in chapter 2, is the causal effect of an information about or a signal sent out by an individual on how this individual is treated by another individual. That is, discrimination refers to behavior, whereas stereotypes refer to knowledge and prejudices refer to evaluation. This distinction is also known as tripartite conceptualization of category based reactions with stereotypes as cognitive, prejudice as affective, and discrimination as behavioral component, respectively (Fiske, 1998, p. 357). Of course, discrimination is not simply the behavioral manifestation of prejudice or stereotypes and, hence, neither equals stereotyping nor applied prejudice (J. M. Jones, 1997).

People with negative prejudices or negative stereotypes do not necessarily engage in discriminatory behavior toward members of the target group—be it on rational grounds or because they follow norms (Gaertner & Dovidio, 1986; LaPiere, 1934; Merton, 1949). On the other hand, rational-calculating or norm-following behavior might also cause people to discriminate against members of a particular group even though they do not hold negative stereotypes about or have negative prejudices against the same group—see, for example, Merton (1949)'s classic typology on this or the contributions to the institutional discrimination literature (e.g., J. R. Feagin & Booher Feagin, 1986; Gomolla, 2016; Gomolla & Radtke, 2010).

5.2 How (Not) to Measure Stereotypes

5.2.1 Explicit Versus Implicit Measures of Stereotypes

Probably the most important distinction between various measures of stereotypes is the one between explicit measures, or measures of explicit *beliefs*, and implicit measures, or measures of implicit *associations*. Implicit measures of stereotypes and attitudes are relatively recent tools that have created a lot of attention among social scientists. These measures—e.g., priming methods (e.g., Fazio et al., 1995) or the implicit association test (IAT) (Greenwald & Banaji, 1995) and related tasks—"rely on processes that are uncontrolled, unintentional, autonomous, goal-independent, purely-stimulus-driven, unconscious, efficient, or fast" (De Houwer & Moors, 2007, p. 192)—or at least "less controllable by respondents" than explicit measures (Fazio & Olson, 2003, p. 636; Gawronski, 2009).

Therefore, it has been suggested that implicit measures have two major advantages over explicit measures: Firstly, implicit measures should be less sensitive to social desirability bias (Fazio et al., 1995, p. 1022; Greenwald et al., 1998, p. 1465). In an "era of contested prejudice" (Lucas, 2008), respondents might shy away from honestly reporting their stereotypes to not violate personal or societal norms (also see, e.g., Gaertner & Dovidio, 1986). Secondly, implicit measures could allow for a more accurate measure of stereotypes, since respondents might lack introspective access to implicitly stored associations and, hence, would simply be unable to accurately report all aspects of a stereotype explicitly (Hofmann & Wilson, 2010).

Not only have these supposed advantages been called into question (Gawronski, 2009; Gawronski et al., 2007), there are also at least two major advantages of explicit measures that made them our method of choice to assess teachers' stereotypes in the NEPS: Firstly, they are very easy to implement, as the researcher only has to ask one or more questions and the respondent answers in more or less closed form. Secondly, explicit measures usually can be implemented in a paper-pencil survey questionnaire like they are used in the NEPS and filled in by the respondents without assistance[26]. Hence, they do not require additional data collection and are, thus, more cost effective in a large scale survey such as the NEPS.

26 While there are paper pencil versions of the IAT and other implicit measures (see, e.g., Vargas et al., 2007, for a review), they are all still much more complex than explicit closed-ended questions, where one item can be enough to assess the stereotype dimension of interest.

5.2.2 A Brief History of Explicit Measures of Stereotypes

Explicit measures of stereotypes have a long history in social science research. Katz and Braly (1933) measured stereotypes by using an *adjective checklist*. This method asks the respondents to select those adjectives they consider to be typical of a particular group of people. The adjective checklist yields estimates of socially shared stereotype contents in the aggregate presumably due to both prevalence and extremity of individual stereotypes. However, on the individual level these measures are less useful, as differences between groups on a particular dimension cannot be quantified beyond the dichotomy 'mentioned-not mentioned'. For example, asking teachers whether they think particular groups of students are "hard-working" or not does not allow teachers to rank several groups according to *how* hard-working they supposedly are. The only feasible solution would be to present multiple items that qualify the adjective of interest—e.g., "somewhat hard working", "rather hard-working", "extremely hard-working", and so on. However, even this procedure does not yield measures that allow to assess the accuracy of teachers' beliefs. Also, neither is it a very efficient way to investigate beliefs, nor will it yield data that can readily be used for data analyses that aim for testing hypotheses derived from theories of discrimination, such as statistical discrimination theory.

Percentage estimates or *scale ratings* are usually used to assess the prevalence of a stereotype but also allow for a more nuanced rating of groups. Percentage estimates ask the respondent to estimate the proportion of people from a social group that is characterized by an attribute or engages in a particular behavior (see, e.g., B. Park & Rothbart, 1982). Brigham (1971), for example, used percentage ratings to assess how prevalent respondents believe a particular characteristic or behavior is among a particular group of people and to quantify the deviation of individuals from the average respondent in the sample. This way Brigham (1971) seeks to identify unjustified generalizations, precisely what he defines as a stereotype. The method of percentage estimates could be used to assess the beliefs of teachers about the percentage of students from different ethnic and social groups that show a particular skill or ability or successfully complete a particular track. Such a measure would allow to rank different groups of students and—as long as the true prevalence is known, e.g., how large the proportion of students with a Turkish background is that successfully completes *Gymnasium*—to assess the accuracy of teachers' beliefs.

Similarly, in scale ratings respondents either rate the likelihood or how typical it is that a member of a social group features a particular attribute or engages in a particular behavior. These ratings can be very similar to percentage estimates and, hence, sometimes the two are treated interchangeably (e.g., D. J. Schneider, 2004, p. 40).

However, how similar scale ratings are to percentage estimates obviously depends on the scale and the items used. Items that describe behavior and then ask how much the respondent agrees that this behavior occurs (e.g., Glick & Fiske, 1996) are not very useful for purposes of assessing the accuracy of teachers' beliefs—in particular with regard to group differences. Combining questions that ask for a clearly quantifiable characteristic with scales that represent or mimic the corresponding units seems to be a much more useful approach.

The *stereotype differential technique* (Gardner, 1973) builds on the methodology of the semantic differential (Osgood et al., 1957) to assess respondents' stereotypes. Respondents rate social groups on a bipolar scale—usually a 7-point scale—with endpoints labeled with opposing adjectives or traits. Socially shared or cultural stereotypes are defined through a significant deviation of the sample mean from the scale's midpoint and through the standard deviation in the sample, where a smaller variation means more consensus. An individual's stereotype score could be obtained by summing up an individual's ratings on those dimensions identified as being part of the cultural stereotype (Gardner, 1973, p. 141).

Yet another way of measuring stereotypes is the *diagnostic ratio*, suggested by McCauley and Stitt (1978). Applying a Bayesian logic, the authors argue that former methods ignore baseline probabilities and suggest that a valid measure of stereotypes has to relate group specific estimates of the prevalence of a particular characteristics or a particular behavior to estimates how prevalent the same characteristic of behavior is among all people.

Methods that focus on the distribution of a particular characteristic or behavior among members of a group of people are the so called *histogram* or *distribution task* (B. Park & Judd, 1990; Wyer et al., 2002) and range task (B. Park & Judd, 1990). While the former—drawing a histogram or distribution of a characteristic within a social group—seems to be too much of a burden for some respondents (B. Park & Judd, 1990, p. 175), the range task is considered a fairly easy to understand measure that yields reliable estimates of both stereotypicality and dispersion (Correll et al., 2010, p. 53).

However, none of these methods—stereotype differential technique, diagnostic ratio, histogram task, distribution task, or range task—yields informative quantitative individual level data that is easy to collect through a concise instrument in paper-pencil self-administered questionnaires. At the NEPS we concluded that some kind of simple and straightforward rating scale approach would be the most promising way to end up with a quantitative measure that yields within- and between-teacher variation that could be used in a statistical model. How we developed and improved our measure is described in the next section.

5.3 Development of an Item Battery to Assess Teacher's Stereotypes[27]

Since NEPS uses paper-pencil self-administered questionnaires for educators and teachers at all stages, implicit measures were unfeasible to implement and we turned to explicit measures instead. The first assessment of teachers' stereotypes about the performance of different groups of students takes place in the fourth wave of Starting Cohort 2 ("Kindergarten and Elementary School"). This wave focuses on second grade students and features interviews with their teachers and parents. We implemented measures of stereotypes in this cohort and at this early stage of the academic career since effects of stereotypes on academic performance are reported to be strongest among the youngest pupils (Jussim, Robustelli, et al., 2009, p. 360). Measures of child competencies undertaken at later times may, thus, be influenced by teachers' stereotypes.

Because of the limited scope of the questionnaires and our interest in several groups of students, we had to restrict our measure to one key dimension. Theory suggests that the single most important belief for teachers' judgments in grading, ability grouping, and track recommendations should be the performance of a student or, for that matter, the average performance of the group the student is categorized in by the teacher. This is backed up by empirical studies that find individual test scores to be the best predictor of grades and track recommendations at the end of elementary school in Germany (see, e.g., Bos, Tarelli, et al., 2012). Also, teachers explicitly name competencies and educational achievement or performance as the most important determinant for their decision which track to recommend (Stahl, 2007). Surprisingly, we neither found an explicit measure of teachers' stereotypes about average group competencies readily available, nor did we find a measure that could have served as a starting point. Hence, we developed our own explicit measure of teachers' beliefs about the average competencies of students from various social and ethnic groups. On this way, we had to answer the following questions:

Which groups? At the NEPS, we decided to ask teachers to report their stereotypes about those groups on which researchers in German education have—for various reasons—recently focused (for reviews see Kristen et al., 2011; Stocké et al., 2011). These groups are: Girls, boys, students with lower, middle, and upper social class

27 The development of the instrument introduced in this section was a collaborative endeavor of pillar 3 and pillar 4 of the National Educational Panel Study (NEPS) at the University of Bamberg, Germany, now at the Leibniz-Institute for Educational Trajectories. Melanie Olczyk (at the time pillar 4) and I (at the time pillar 3) were mainly responsible for developing and (pre)testing the instrument. Also see Wenz et al. (2016) on which this chapter is based.

background, students of Turkish and Russian origin, as well as immigrants and ethnic majority students in general. This includes the groups I focus on in this dissertation.

Which stereotypes exactly should we assess? We decided to assess teachers' stereotypes regarding mathematical and reading competencies because math and German are the two major subjects in German elementary school whose grades in several states are explicit legal determinants of the track recommendation at the end of elementary school (Helbig & Nikolai, 2015). Also, math and reading competencies are assessed in wave 4 of starting cohort 2 in the NEPS, which enables researchers to directly investigate the accuracy of teachers' stereotypes.

How to ask for stereotypes? The introduction serves the purpose of a cover story and is supposed to reduce social desirability bias by turning teachers' attention to the NEPS competence tests—instead of just asking for general or innate abilities or competencies of groups of students. Therefore, the introduction for the item battery reads as follows (see also Figure 1):

> In the NEPS study "Educational trajectories in Germany" the competencies of children are assessed in different domains. What do you think how second graders from various groups will perform in mathematics [reading]?

Through this introduction we intended to direct the teacher's attention away from the assessment of stereotypes and to their diagnostic competence as experts about educational achievement and competencies of different groups of students. Put differently, we reckoned that the framing of the question would be crucial for a high response rate as well as for keeping the social desirability bias as low as possible. Also, we wanted to ask in a general way that would allow teachers to report whatever they think of first when thinking of the competencies of different groups.

In which order should we ask for groups? The initial idea about the item order—unfortunately not mentioned in Wenz et al. (2016)—was twofold: First, the item battery should start with girls and boys to provide teachers a gentle start into a series of questions that some might find unpleasent to think about and, thus, answer. Secondly, that the generic group of "immigrants" appears before the groups of students with Turkish and Russian immigrant background was a deliberate choice that aims at avoiding an assimilation effect as it typically occurs when specific items precede general items that tap the same domain (Schwarz et al., 1991). That the items on the different social classes appear before the items on immigrants and that ethnic majority students are assessed at the end had no particular reason.

Which response scale should we use? For the response scale we had three major criteria: First, we wanted to allow teachers to express the belief that a particular group's competencies are average and, therefore, decide that the scale should have a mid-

In the NEPS study "Educational trajectories in Germany" the competencies of children are assessed in different subjects. What do you think how 2nd graders from various groups will perform in mathematics [reading]?

Compared to the mathematics [reading] performance of 2nd graders in general…

The further left you tick, the worse the group will perform according to your estimate, the further right you tick, the better. Please tick one square each line.

		perform very poorly									perform very well
		0	1	2	3	4	5	6	7	8	9	10
a)	… girls	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
b)	… boys	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
c)	… children from lower social strata	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
d)	… children from middle social strata	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
e)	… children from higher social strata	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
f)	… children of immigrant origin	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
g)	… children of Turkish origin	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
h)	… children of Russian origin	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
i)	… majority	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Figure 5.1: First version of the instrument to measure teachers' stereotypes in the NEPS. Figure including translation adopted from Wenz et al. (2016)

point. Secondly, we wanted to avoid confusion with the German grading scale that ranges from 1 for "very good" to 6 for "insufficient" or "failed", respectively. Thus, the labels of the response scale should be sufficiently distinct from the German grading scale. Thirdly, we wanted to allow teachers to express finely nuanced beliefs through enough points on our response scale. Taken together, these criteria led us to pick an 11-point scale—instead of 9-, 7-, or 5-point scale.

Based on these considerations we developed an item battery featuring two items for each of the nine groups, asking about the average level of math and reading competencies, respectively. Figure 5.1 shows the first version of our instrument (see Appendix C for the original version in German language).

5.3.1 Developing the Instrument and Assessing its Validity Through Cognitive Interviews

This first version (see figure 5.1) was modified after feedback from both colleagues and teachers with whom we conducted cognitive pretests (Desimone & Floch, 2004;

Miller et al., 2014; Schwarz & Sudman, 1996; Willis, 2004). We evaluated our instrument through structured interviews with five teachers, recruited in the region of Bamberg, Germany. The interviewed teachers taught at least mathematics or German. Two teachers were working in elementary schools, three in secondary schools. The recruitment of these teachers was realized through social contacts within the NEPS project. We probed participants retrospectively by asking them immediately after each question they had answered to tell us about, for example, how they understood the questions and their understanding of key terms used in the question. We decided against the think aloud technique, to not disturb the thought process that respondents go through when answering our items and keep the respondents' burden as low as possible (Collins, 2003; Willis, 2004, pp. 52–57). Results from the cognitive interviews lead to three major modifications of the first version (see figure 5.2 and Appendix 2):

Lead-in: While in the first version (see figure 5.11 and Appendix 1) it was asked how children attending the second grade perform compared to second graders, in the second version—which was also implemented in the pilot study of the NEPS—we added a concrete reference and asked teachers to report their beliefs "[…] compared to the average". We did this because through the cognitive interviews we learned that teachers almost exclusively referred to students in their current or previous classes. We wanted the question to allow for a broader understanding of it. To test the modification of the lead-in, we recruited new teachers for cognitive pretests and conducted four further interviews. All four teachers were working in elementary schools in or near Bamberg and taught mathematics and German at the time. The results indicate that the modification successfully turned attention away from the own students to the performance of second graders more generally.

Repetition of the question wording: In the revised second version (see figure 5.2) we repeated the key question and separated the different groups to remind the teachers of the task at hand. This was done to assure that teachers use the same anchor of reference for all judgments, and, thus, to avoid unwanted assimilation and contrast effects (Schwarz et al., 1991).

Labels of the response scale: In addition, results of the cognitive interviews suggested that the initial labeling of all numerical values from 0 to 10 on the response scale might have been misleading to some of the teachers who had in mind the German grading scale, which ranges from 1 to 6. Apparently, they ticked the value they had in mind in terms of a grad—e.g., 2 for "good"—ignoring the other values and the endpoint labels. By restricting the labels to values 0, 5, and 10 we aimed at decreasing the likelihood of such misunderstandings but still allow teachers to successfully navigate the scale.

In the NEPS study "Educational trajectories in Germany" the competencies of children attending the 2nd grade are assessed in different domains.

What do you think how 2nd graders from the following groups will perform compared to the average in the domain mathematics [reading]?

The further left you tick, the worse the group will perform according to your estimate, the further right you tick, the better. Please tick one square each line.

		very poorly 0					5					very well 10
a)	Girls	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
b)	Boys	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

And how will the following groups perform compared to the average?

		very poorly 0					5					very well 10
c)	Children from lower social strata	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
d)	Children from middle social strata	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
e)	Children from higher social strata	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

And how will the following groups perform compared to the average?

		very poorly 0					5					very well 10
f)	Children of immigrant origin	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
g)	Children of Turkish origin	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
h)	Children of Russian origin	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
i)	Majority	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Figure 5.2: Second version of the instrument to measure teachers' stereotypes in the NEPS. Figure including translation adopted from Wenz et al. (2016)

5.3.2 The Final Version

In the final version of the instrument (see figure 5.3 and Appendix 3)—which was implemented in the main study and, hence, through which the data for the scientific use files of the NEPS is collected—three further modifications were undertaken based on discussions with colleagues:

Lead-in: We modified the lead-in to what might be translated to "What do you think how second graders from the following groups will perform compared to all second graders in Germany in the domain mathematics [reading]?" Hence, while in the second version "the average" is the reference, in the third and final version a more precise description of the reference, namely "all 2nd graders in Germany" is used.

Labels of the response scale: For the final version, we changed the labels of the response scale from perform "very poorly" and perform "very well" into perform "far below average" and "far above average". The aim was to make the scale more relative, stress the reference group ("all second graders in Germany"), and, in consequence, to achieve a less skewed distribution as well as more variance. This change also serves the purpose of allowing accuracy of beliefs to be assessed against the performance of a group relative to the sample mean, now that the midpoint of the scale has a meaning in terms of the actual competence distribution.

Order of social groups: Finally, questions referring to students of different sex and different social origin were swapped. In consequence, the query now starts with children from lower social strata instead of girls. The aim of this approach was to avoid a pattern where respondents contrast their responses within the social groups, for example, by referring to girls when estimating the performance of boys—rather than referring to all children attending the second grade. This was one of the few decisions we made during the process of developing this measure that I was not happy with, since I did and do not see how asking about girls and boys after asking about students from different social strata will make respondents less likely to refer to girls when reporting their beliefs about the performance of boys. In fact, I think that starting with girls and boys is more gentle than starting with students from different social strata that might decrease response rates or otherwise affect the validity of teachers' answers[28].

28 Whether this is an effect of their pedagogical training or more due to self-selection, teachers in Germany seem to be rather cautious when it comes to answering questions that generalize over groups of students—at least according to my experience at the NEPS.

More results from cognitive interviews

Taken together, we conducted nine cognitive interviews to pretest the instrument. In addition to the results reported above, the interviews show that teachers share a common understanding of key terms used in the questions. Also, their understanding of both questions and key terms is similar to our understanding. Teachers understand that the questions aim at their personal assessment. Two of the nine teachers explicitly reckoned that the questions aim at their stereotypes about certain groups and the stereotypes' potential influence on the academic success of students from these groups. However, there was no evidence that teachers would change their responses because of that.

Almost all interviewed teachers define social strata mainly through parental income or parental education or a combination of the two. In addition, some refer to the occupational status of the parents as well as to the learning environment and support at home. All in all, the teachers tend to have a similar understanding of the various social strata. Only one teacher had problems classifying different social strata. According to the interviewed teachers, lower social strata are characterized by living on welfare or a relatively low household income or a less beneficial learning environment at home. The middle social strata are associated with higher income than the lower strata. The higher social strata are associated first and foremost with high education—for example, a high rate people holding tertiary degrees—and, therefore, also also with higher financial resources.

Seven of the nine teachers provided definitions of persons of immigrant origin. Again, the results show that teachers largely agree: Almost all teachers referred to individuals who were born in a foreign country themselves or have at least one parent born abroad. Only one teacher restricted the definition to first generation immigrants. As with social strata, teachers also explain their understanding of the term by mentioning educational outcomes as indicators of an immigrant background: Six of the seven teachers mention language competence and language use in the home environment as criterion. Furthermore, when estimating the performance of children of Russian origin, all interviewed teachers consider children from today's Russia as well as children with parents born in the Soviet Union and its successor states.

With regard to reading and mathematics competencies, teachers show a rather similar understanding of these terms. Finally, there is no evidence that teachers were led astray by the fact that the value labels for the endpoints of the scale range over more than one box.

In the NEPS study "Educational trajectories in Germany" the competencies of children attending the 2nd grade are assessed in different domains.

What do you think how 2nd graders from the following groups will perform compared to all 2nd graders in Germany in mathematics [reading]?

The further left you tick, the worse the group will perform according to your estimate, the further right you tick, the better. Please tick one square each line.

		far below average 0					5					far above average 10
a)	Children from lower social strata	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
b)	Children from middle social strata	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
c)	Children from higher social strata	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

And how will 2nd graders from the following groups perform compared to all 2nd graders in Germany?

		far below average 0					5					far above average 10
d)	Girls	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
e)	Boys	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

And how will 2nd graders from the following groups perform compared to all 2nd graders in Germany?

		far below average 0					5					far above average 10
f)	Children of immigrant origin	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
g)	Children of Turkish origin	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
h)	Children of Russian origin	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
i)	Majority	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Figure 5.3: Final version of the instrument to measure teachers' stereotypes in the NEPS. Figure including translation adopted from Wenz et al. (2016)

5.4 Data and Analytic Strategy[29]

In this section I briefly describe the data I use and the analytical strategy I adopt to answer the research questions raised in the beginning of the chapter: First, can we validly measure teachers' stereotypes about different groups of students? If so, do teachers' stereotypes, secondly, differ between groups? And, thirdly, are teachers' stereotypes accurate or more or less biased and, if so, to the disadvantage of which groups of students? Note that these questions are also investigated in Wenz et al. (2016), who argue that the latter questions are informative with regard to the quality and validity of the new instrument. I will return to the arguments below in section 5.4.2

5.4.1 Data

As in Wenz et al. (2016), the analyses below are based on data of the fourth wave of the pilot study in the NEPS Kindergarten cohort. At the time the survey was conducted, the children attended the second grade. The main aim of the NEPS pilot studies is to guarantee smooth main studies—for example by testing instruments and fieldwork. Just like the main studies, the corresponding pilot studies are conceptualized as panel studies. The sampling procedures of main and pilot studies are essentially equivalent. However, the pilot studies feature smaller samples and are conducted only in selected federal states of Germany. The sample for the pilot study in the Kindergarten cohort was drawn on four states: Bavaria, Thuringia, North Rhine-Westphalia, and Hamburg. In total, I can draw on 52 teacher interviews. Note that data from NEPS pilot studies is not released to the scientific community.

All quantitative results reported in this section refer to the second version of the instrument (see figure 5.2). This version was implemented only in the pilot. Since the final version of the instrument (see figure 5.3) differs only slightly, I expect conclusions to be substantively the same in replications using the final version of the instrument as long as data from the same federal states are used. Replications using all data from the main study might differ due to differences between teachers in different federal states.

29 Syntax for all quantitative analyses in this chapter is available at https://osf.io/dqtkg/. However, the data from the NEPS pilot study is unfortunately not available for replication purposes.

5.4.2 Analytic Strategy

While I will largely reproduce the analyses of Wenz et al. (2016), I shift the emphasis slightly towards answering two of the three research questions asked in the beginning of this chapter: Do teachers' stereotypes regarding mathematical and reading competencies of different groups of students differ between groups? And: Are these stereotypes accurate or are they biased and, if so, to the (dis)advantage of which groups? In answering both questions, I go beyond Wenz et al. (2016) regarding both theorizing what to expect and analyzing what teachers' told us.

I stick to the perspective in Wenz et al. (2016) that both questions relate to the quality and validity of the instrument—and, thus, to question one—and check the instrument for the following desirable properties: (i) variation between groups as a consequence of variation within teachers, (ii) variation between teachers, (iii) validity of the measure, and for the rather undesirable property of (iv) missing values. While examining missing values as well as variation within and between teachers is straightforward, validating the measure is less so. With regard to different forms of validity—content, criterion, and construct validity—I argue that content validity is satisfied by the question wording that rather explicitly asks for what has been defined as stereotype above. Remember that the cognitive interviews provided evidence that teachers understand the questions as intended, which, in turn, provides evidence in favor of the content validity of the instrument (Miller et al., 2014, p. 3; Haynes et al., 1995; Willis, 2004).

To assess criterion validity and construct validity—of which the latter is often seen as the most general, overarching form of validity that encompasses criterion and content validity as special cases (e.g., Haynes et al., 1995)—I perform the same quantitative tests I have performed in Wenz et al. (2016). I also stick to the line of argument laid out in Wenz et al. (2016) and suggest that if the instrument is a valid measure of teachers' stereotypes, mean differences between groups, corresponding effect sizes, and correlations should follow theoretical expectations and previous research as summarized in section 5.4.3 below.

Assessing accuracy

There are different ways of defining and assessing accuracy (e.g., Jussim, 2012, pp. 170–194; Jussim et al., 2015). However, an in-depth discussion of different conceptualizations and methods is beyond the scope of this chapter. Instead I briefly present my understanding of accuracy and bias as well as the methods I will use below to assess accuracy in teachers' beliefs as measured using the new instrument.

My conceptualization distinguishes between a dichotomous understanding of ac-

curacy and a dimensional understanding: From a dichotomous perspective, a belief is *accurate* if and only if it is correct or true and, thus, not inaccurate. Put differently, as soon as a belief is not true or correct, it is inaccurate. Obviously, this perspective applies to questions to which there are only two possible, one correct, one incorrect answer or at least answers that can be meaningfully dichotomized. However, dichotomization typically involves loss of information and, thus, I also acknowledge that beliefs about categorical or continuous outcomes require a more nuanced language that allows beliefs about groups to range "from completely accurate to completely inaccurate" (Jussim et al., 1995, p. 16). This implies that beliefs, assessed against such an accuracy-inaccuracy dimension (Jussim et al., 1995, pp. 16–17) can be more or less accurate. Not only is this useful for comparing the accuracy of beliefs of different people or groups of people but also because, strictly speaking, some beliefs can never be exactly right. This holds for all beliefs about averages or other scalar values of continuously distributed group characteristics, such as body height of men and women or competencies of different groups of students, since the probability for one exact value along a continuous distribution is zero. This issue is sometimes addressed by defining a range around the true value as constituting an accurate belief (Jussim et al., 2015, p. 492). I do not find such dichotomization strategies particularly helpful—first, because of a loss of information; secondly and more importantly, because defending a particular range is virtually impossible using scientific methods.

Note that while I have used and will use *bias* as an antonym to accuracy and thus *inaccurate* and *biased* interchangeably, I typically use inaccurate when I simply mean false or incorrect in an absolute sense but biased when I refer to the under- or overestimation of a group relative to another group.

Now which strategies of assessing accuracy in teachers' stereotypes about groups of students will I (not) use below? The developed instrument does not allow to assess *accuracy in an absolute sense*, that is, it is not possible to simply take the belief about a group's average performance and compare it to a criterion taken from real data—be it the NEPS competence test or any other competence test or published results. The reason is that both the instrument assessing teachers' stereotypes as well as contemporary competence tests in education have no straightforward absolute interpretation but only make sense in relative terms.

A *relative comparison that focuses on one group only* would be possible by assessing the accuracy of beliefs against the performance of a group relative to the sample mean. However, the data used here do not allow for such an interpretation, because of the absolute response scale used in the second version of the instrument. Such comparisons will be possible with data from the NEPS scientific use files, since the final ver-

sion of the instrument (see figure 5.3) used a scale labeled ranging from "far below average" to "far above average". The midpoint (5) of this scale can then be interpreted as coinciding with the sample average of the actual competence distribution.

Other relative assessments of stereotype accuracy have to be built on *group comparisons* involving at least two groups. One possibility is to calculate *effect sizes of group differences* that—under certain assumptions—allow to directly compare group differences as perceived by the teachers with group differences as reported in published studies[30]. How exactly I calculate the effect sizes is explained further below.

Using both effect sizes and simple mean differences I will also look at *comparisons in relation to other reference groups and across domains*. Comparisons in relation to reference groups may involve comparing relative distances of, for example, two groups of immigrants to the ethnic majority. Across domains, too, relative distances can inform the assessment of accuracy. Take, for example, boys and girls whose performance differences in mathematics and their performance differences in reading might be of similar size but estimated to of different size. Note that these strategies follow a difference-in-differences logic, that is, bias is only detected if a group is overproportionally over- or underestimated by the teachers, not if all groups are over- or underestimated by the same factor.

A rather simple method of assessing the accuracy of teachers' stereotypes is to investigate whether teachers get the *relative ranking of groups* right. For a single teacher and two groups, I and J, there are three possible outcomes: (1) The groups are believed to perform equally; (2) group I is believed to perform better than group J, or (3) vice versa. Given a sample of teachers, bias in teachers' beliefs can then be quantified by calculating the proportion of teachers that correctly assess the groups' relative positions.

Calculating effect sizes

To enable comparisons of mean differences free of their original measurement units, Cohen (1977, p. 20) suggested to use the effect size d, defined as the mean difference of two groups or populations divided by the standard deviation of either group or population that are assumed equal. To estimate d from sample data I follow Cohen (1977, pp. 66–67) and calculate

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s^*}, \tag{5.1}$$

30 Data from the NEPS competence tests were not available for such calculations when the analyses for this chapter were conducted. However, this will be possible with data from the NEPS scientific use files.

where

$$s^* = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{5.2}$$

and where s_1^2 and s_2^2 are the unbiased sample variances of x_1 and x_2 that are the beliefs of teachers about average performances of groups 1 and 2, respectively. Thus, $\bar{x}_1$ and $\bar{x}_2$ are the means of the beliefs about the performances of groups 1 and 2, respectively, over all teachers. Note that this strategy assumes that s_1^2 and s_2^2 are valid proxies for the average dispersion of groups 1 and 2, respectively, as perceived by the teachers.

This assumption has to be made since the instrument does not directly measure the perceived variability or variance on the level of the individual teacher. In principle, this could have been accomplished by a paper-pencil self-administered questionnaire implementing some kind of distribution task or range task (see section 5.2.2). However, the limited space in the NEPS questionnaires and the risk of decreasing response rates on the level of teachers and educators lead us to decide against a more elaborated assessment. Therefore, all we know are the estimated group averages in test performance—their variance between teachers is our best estimate for perceived variance within groups.

Correlations

Correlations between estimated average group competencies will also serve as an indicator of validity if they follow theoretically explainable patterns as outlined in section 5.4.3 below.

5.4.3 Theory Driven Validation and Expectations

In the remainder of this section, I will argue that answering the two major substantive research questions of this chapter and the validation of the instrument we developed at the NEPS—that is, question number one—pose related problems, connected by the mechanisms that bring about accuracy and bias in stereotypical thinking. Generally speaking, it seems reasonable to expect that a valid explicit measure of stereotypical beliefs of teachers yields similar results as other, different—but nevertheless comparable—instruments that aim for measuring the same or similar constructs. I am not aware of any studies that assess the accuracy of beliefs about groups of students. The closest and most relevant literature in this regard is research on teachers' beliefs and expectations as self-fulfilling prophecies (Jussim, 1989; Jussim et al., 1996; Jussim & Harber, 2005; Jussim, Robustelli, et al., 2009; Lorenz et al., 2016).

Studies from this line of research suggest that, by and large, teachers' beliefs and expectations—typically about individual students, though—are fairly accurate and not biased heavily by social, racial, or ethnic criteria (see, e.g., Jussim & Harber, 2005; Jussim, Robustelli, et al., 2009, for reviews). Certainly, there is evidence for some bias against particular groups of students in the international literature (e.g., Campbell, 2015; Ready & Wright, 2010) as well as in the German literature (e.g., Lorenz et al., 2016) and I will return to these biases below.

In addition to the empirical studies cited above, there are different *theoretical* considerations that can inform expectations about the accuracy and the biases in teachers' beliefs. For decades research in social psychology has highlighted the biases and inaccuracies of stereotypes—often by merely defining stereotypes as incorrect (see section 5.1). Sometimes by empirically investigating and stressing the biased nature of stereotypes defined in such a way—e.g., "as an exceptionless generalization about the target group (e.g., 'All Asians are smart')" (Jussim et al., 1995, p. 6)—that empirical investigations could only yield the conclusion that stereotypes are biased. Sometimes by confusing the stereotype itself with the outcome of a process of stereotyping (Jussim, Cain, et al., 2009, p. 215) that almost inevitably results in individual discrimination (Aigner & Cain, 1977; England & Lewin, 1989) which, in turn, constitutes a biased view of an individual. Remarkably, the confusion of stereotypes with the process of stereotyping—that is, confusing beliefs about groups with the judgment or treatment of individual group members—is also present in major contributions to the research on stereotype accuracy (Jussim et al., 1996; Jussim & Harber, 2005; Jussim, Robustelli, et al., 2009; but cf. Jussim et al., 2015, who acknowledge the confusion). Only rarely have biases in stereotypes been investigated and, if so, research typically shows that these biases tend to be moderate in size (see, e.g., Jussim, Cain, et al., 2009; Jussim et al., 2015; C. Ryan, 2002, for reviews; Diekman et al., 2002).

Not only has more recent research shown that stereotypes—understood simply as beliefs about the characteristics, attributes, or behaviors of a group of people—are not always inaccurate (Jussim, Cain, et al., 2009; Jussim et al., 2015; C. Ryan, 2002), there is convincing evidence that processes of stereotype formation, categorization, and stereotype application are deeply rooted in the biology of the human brain (Fishbein, 2002, pp. 39–82; Fiske, 2000; Caporael, 1997). It seems highly unlikely that cognitive processes that arguably have been key to the survival of individuals and groups in the evolution of the human race should be built on largely inaccurate beliefs producing largely inaccurate judgments, expectations, or maladaptive behavior. However, note that such an explanation does not rule out particular forms of biases—especially in terms of affect and emotion, that is prejudice—that might, indeed, also increase

reproductive fitness. All this evolutionary perceptive suggests is that beliefs and the related mechanisms should not be grotesquely mistaken or flawed.

In addition to the arguments for why stereotypes in general should not be too inaccurate, there are good reasons to believe that teachers' stereotypes about groups of students in particular should be fairly accurate. Teachers are experts in teaching students from different social and ethnic groups as well students of different gender and can, therefore, be expected to be knowledgeable with regard to the competencies of different groups of students. In fact, experts' judgments are sometimes used as a criterion against which the accuracy of stereotypes is assessed (Judd & Park, 1993, p. 114).

However, convincing theoretical reasons for biases in teachers' beliefs remain and whenever teachers' stereotypes are not accurate—that is biased—these biases should show patterns of ingroup favoritism or outgroup derogation (Tajfel, 1982; Tajfel & Turner, 1986), respectively, if the measure is a valid measure. That is, teachers' assessments should show a bias in favor of groups they belong to, and a negative bias towards those they do not belong to. Since teachers in German elementary schools are overwhelmingly female, belong to the middle or upper (middle) class, and are of German ethnicity without immigration background, I expect biases—if any—to the disadvantage of boys, students from lower class families, as well as immigrants in general, and different groups of immigrants in particular.

Another well replicated phenomenon in intergroup perception is outgroup homogeneity, which means that—under certain conditions—members of outgroups tend to be perceived as more similar to one another than they really are (Judd & Park, 1993; Judd et al., 1991; C. S. Ryan & Bogart, 2001) and more similar than members of ingroups (Brown & Wootton-Millward, 1993; Judd & Park, 1988, 1993; Judd et al., 1991; B. Park & Rothbart, 1982; C. S. Ryan & Bogart, 2001). Especially members of minority or low status groups tend to be perceived this way (e.g., Fiske, 1993a; Lorenzi-Cioldi et al., 1995; but cf. Brauer, 2001).

I suggest that the outgroup homogeneity effect should also hold on the group level: minority outgroups that can easily be categorized into one superordinate minority outgroup (Ashmore & Longo, 1995; Gaertner et al., 1993; González & Brown, 2003; B. Park & Rothbart, 1982) should be perceived as more similar to one another than they actually are. In particular, I expect different groups of immigrants, Turks and Russians, to be perceived more similar than they actually are, as they are easily categorized into a superordinate group of immigrants. Also, teachers' beliefs about groups that can be categorized into one superordinate group or groups that are otherwise perceived to be similar should correlate positively. Therefore—and even though the

chosen item order should reduce a possible part-whole assimilation effect (Schwarz et al., 1991)—, I expect positive correlations between teachers' stereotypes for immigrants and Turks, immigrants and Russians, as well as Turks and Russians.

In contrast, I expect low and insignificant correlations between unrelated groups such as girls or boys on the one hand and different groups of immigrants on the other hand. Especially because they are likely to serve as standards of comparison for each other, teachers' stereotypes about the competencies of boys and girls should correlate positively again. The same logic should apply to the different social classes. Thus, I also expect positive and significant correlations among teachers' stereotypes for the three groups. However, how exactly teachers understand the terms lower, middle, and upper class is less clear than it is for boys and girls. Also, the correlations might in part be due to item order and, thus, items that are further apart can be expected to correlate less than those that are closer together (Schwarz et al., 1991; Weijters et al., 2009). However, predicting patterns of item intercorrelations in more detail seems difficult for theoretical reasons, because how exactly teachers will categorize groups into superordinate groups or how they will use groups as reference for one another is not clear ex-ante, also because the fixed item order could and, indeed, can be expected to influence the process of categorizing and referring to groups.

5.5 Quantitative Results

From the theories discussed in chapter 3 it should be clear that the relation between stereotypes and discrimination is everything but straightforward. This is true for both individual and group discrimination. However, it is worth reminding ourselves which major patterns we should look for and why: When stereotypes are introduced as explanation for discrimination as an individual level causal effect, they will usually have to vary between groups to make such an explanation work. If stereotypes are biased to the (dis-)advantage of certain groups or if other conditions are present, they may also explain group discrimination—even if they do not vary between target groups. For more on the role of stereotypes in explaining discrimination, see section 3.1.2 in particular.

5.5.1 Within Teacher Variation

Figure 5.4 summarizes between group variations as mean differences between teachers' stereotypes of the competencies of different groups. Teachers' stereotypes vary considerably between groups for both math (left panel) and reading (right panel).

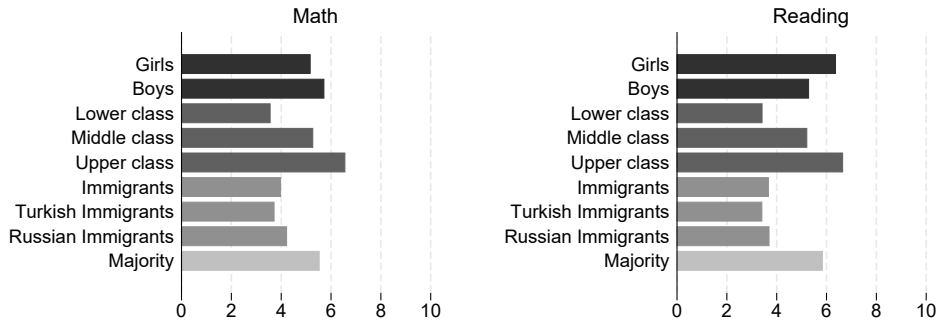


Figure 5.4: Means of teachers' estimation of students' results in NEPS competence tests for math (left panel) and reading (right panel). Groups (number of observations by stereotype in parentheses) from top to bottom: girls (math: 49/reading: 50), boys (49/50), lower class (45/46), middle class (45/48), upper class (45/48), immigrants (40/42), Turkish immigrants (35/37), Russian immigrants (37/39), majority (40/42).

As math and reading competencies are empirically strongly correlated (Rindermann, 2007), it is not surprising that the overall patterns look very similar. However, there are systematic differences with regard to gender, which is the category I start my discussion of results with.

Gender

While teachers believe that boys outperform girls in math (mean difference: $-.55, p < .05$), the opposite is true for reading (mean difference: $+1.08, p < .001$). Empirical studies provide strong evidence for this pattern (Bos, Tarelli, et al., 2012; Bos, Wendt, et al., 2012; Mücke & Schründer-Lenzen, 2008). However, the same studies suggest that the teachers are mistaken in estimating the advantage of girls in reading (Cohen's d: $.78$) to be about twice as large as boys' advantage over girls in mathematics (Cohen's d: $-.41$). Nationwide evidence for fourth-graders suggests that the gender gaps in math and reading competencies are of similar size (mean difference in reading: $+8$, $d = .12$, Bos, Tarelli, et al., 2012, pp. 97, 126; mean difference in math: -8, $d = -.13$, Bos, Wendt, et al., 2012, pp. 65, 208)[31]. Using data from a longitudinal study of 26 schools in Berlin, Mücke and Schründer-Lenzen (2008) moreover find that boys' advantage in math is even larger than their disadvantage in reading. Interestingly, the

31 Neither Bos, Wendt, et al. (2012) nor Bos, Tarelli, et al. (2012) report d or any other effect size. I calculated d using the information given on the pages I cite above.

teachers' stereotypes are consistent with the German results in the PISA study (e.g., Prenzel et al., 2013)—a study German media has covered extensively. If and only if the assumption holds that the variances between teachers are valid proxies for the teacher's perception of dispersion, it is also possible to conclude that the differences between girls and boys in both domains are overestimated by the teachers. How accurate teachers are with regard to ranking boys and girls in both domains is investigated below in section 5.5.2.

Social class

Teachers perceive large competence differences between students from different social classes (math: lower - middle: $d = -1.44$, middle - upper: $d = -1.04$, lower - upper: $d = -2.10$; reading: lower - middle: $d = -1.35$, middle - upper: $d = -1.106$, lower - upper: $d = -2.35$; all differences are statistically significant with $p < .001$). Figure 5.4 confirms the similar patterns for math and reading. However and despite the promising results from the cognitive interviews, it is not clear what teachers had in mind exactly when reporting their expectations about the competencies of different social classes and, thus, whether their understanding matches any operational definition used in published studies. Thus, I cannot assess teachers' accuracy as precisely as for their stereotypes on gender differences. The means over all teachers suggest that, on average, teachers correctly rank the three groups: All available studies show that—whatever the exact operational definition—students from lower class families perform worse than those from middle class families who, in turn, perform worse than those from upper class families (e.g., Bos, Tarelli, et al., 2012; Bos, Wendt, et al., 2012; Prenzel et al., 2013; Stanat et al., 2012).

However, effect sizes, calculated based on numbers reported in published studies, suggest that teachers probably overestimate the differences between students from different social classes. Based on Stanat et al. (2012, p. 202), I find that students from upper middle class and upper class families (EGP classes I and II) perform better than students from middle and lower working class families (EGP classes V, VI, VII) with $d = -.79$ in math and $d = -.81$ in reading [32]. The group specific means in Bos, Wendt, et al. (2012, p. 241) and Bos, Tarelli, et al. (2012, p. 185) allow for even more extreme comparisons, such as the difference between students from upper class families (EGP I) and students from low skilled working class families (EGP VII) that yield a difference of $d = -.74$ in math (Bos, Wendt, et al., 2012) and $d = -.80$ in reading (Bos, Tarelli, et

32 Stanat et al. (2012) do not report group specific standard deviations, so I divided the mean group differences by the standard deviation of the test for the whole sample ($SD = 100$).

al., 2012)[33]. Of course, less extreme comparisons that might resemble the comparison of upper class and middle class or middle class and lower class, yield considerably lower effect sizes, which is why I don't report them here. Since all effect sizes based on empirical studies are smaller than the smallest effect size of those quantifying the differences as perceived by the teachers, the conclusion that teachers are biased to the disadvantage of students from lower class families in more generally and to the disadvantage of students from middle class families in comparison with upper class families seems justified.

Immigrant background

On average, teachers also correctly believe that ethnic majority students perform better on average than their peers with an immigrant background in both math and reading. Virtually all empirical studies on the achievement of immigrants and their descendants in the German education system have shown that they perform worse than students without immigrant background in both math and reading (see, e.g., Bos, Tarelli, et al., 2012; Bos, Wendt, et al., 2012; Stanat et al., 2012, for results for fourth graders). Given the extensive media coverage of these studies and of the comparatively low achievement of immigrants in particular, this result is not surprising. However, judging by effect sizes it seems that teachers overestimate the immigrants' disadvantage to the ethnic majority: While the teachers in the NEPS pilot sample see differences of $d = -1.63$ in math and $d = -1.41$ in reading, Stanat et al. (2012, pp. 214–221) report differences of $d = -.31$ in math and $d = -.26$ in reading for children of which only one parent is born abroad and of $d = -.56$ in both math and reading for children of which both parents are born abroad. Group specific means and standard deviations reported in Bos, Wendt, et al. (2012, p. 258), Bos, Tarelli, et al. (2012, p. 199) result in very similar effect sizes for these groups and domains. My own calculations based on the numbers reported in Stanat et al. (2012) combining both groups into one group of immigrants by using weighted means and weighted variances, yield $d = -.46$ for math and $d = -.44$ for reading as effect size of the immigrant disadvantage. Again, these numbers suggest that teachers underestimate students with immigrant background relative to students without immigrant background in both domains.

The results for students of Turkish and Russian origin are similar to those for im-

33 Although group specific standard deviations are reported in both texts, here, too, I divided the mean group differences by the standard deviation of the test for the whole sample ($SD = 62$ in Bos, Wendt, et al., 2012; $SD = 66$ in Bos, Tarelli, et al., 2012), since neither text provides group specific numbers of observation.

migrants in general: On average, teachers are correct in believing that students of Turkish origin perform worse than those of Russian origin in both math (mean difference: $-.45,\ p < .05$) and reading (mean difference: $-.34,\ p = .07$) (for results for both groups of immigrants in both domains, see Stanat et al., 2012, p. 225; for similar results using data from the PISA studies of 2000, 2003, and 2006, see Walter, 2009). That teachers also get the ranking of these two groups of immigrants right on average is much more remarkable than the ranking of immigrants and ethnic majority, since German media has rarely covered comparisons of different groups of immigrants in general or the performance of students with a Russian background in particular. Effect sizes of the differences between ethnic majority students and students of both Turkish and Russian origin suggest that teachers perceive larger differences between the majority and the two groups of immigrants (disadvantage of Turks: $d = -1.72$ for math, $d = -1.62$ for reading; disadvantage of Russians: $d = -1.17$ for math, $d = -1.44$ for reading) than actually exist (disadvantage of Turks: $d = -.93$ for math, $d = -.88$ for reading; disadvantage of Russians: $d = -.35$ for math, $d = -.34$ for reading), as suggested by my calculations of effect sizes based on numbers reported in Stanat et al. (2012, p. 225). Based on these numbers, it seems that teachers underestimate both groups of immigrants relative to their ethnic majority peers—this pattern might be due to an absolute underestimation of both groups, an absolute overestimation of the ethnic majority, or both. As expected, teachers believe that students of Turkish and Russian origin are more similar than they really are: The disadvantage of Turkish students is estimated to be $d = -.36$ in math and $d = -.18$ in reading, while my calculations based on Stanat et al. (2012, p. 225) yield $d = -.57$ in math and $d = -.56$ in reading as differences between these two groups of immigrants.

Should the reader not buy into the assumption necessary to make the effect size comparisons work (see paragraph *Calculating effect sizes* in section 5.4.2 above), it might be of interest to learn that analyses using relative ratios of mean differences only, as conducted in Wenz et al. (2016), by and large yield the same conclusions as based on effect sizes.

5.5.2 Between Teacher Variation

Since the same between group variation may stem from few teachers perceiving large differences or many perceiving small differences, figures 5.5 and 5.6 show the variation within teachers as a difference between two selected groups as rated by the same teacher. For both math and reading the plots show that teachers differ to some degree in their estimates of group differences: Not only are different groups of students

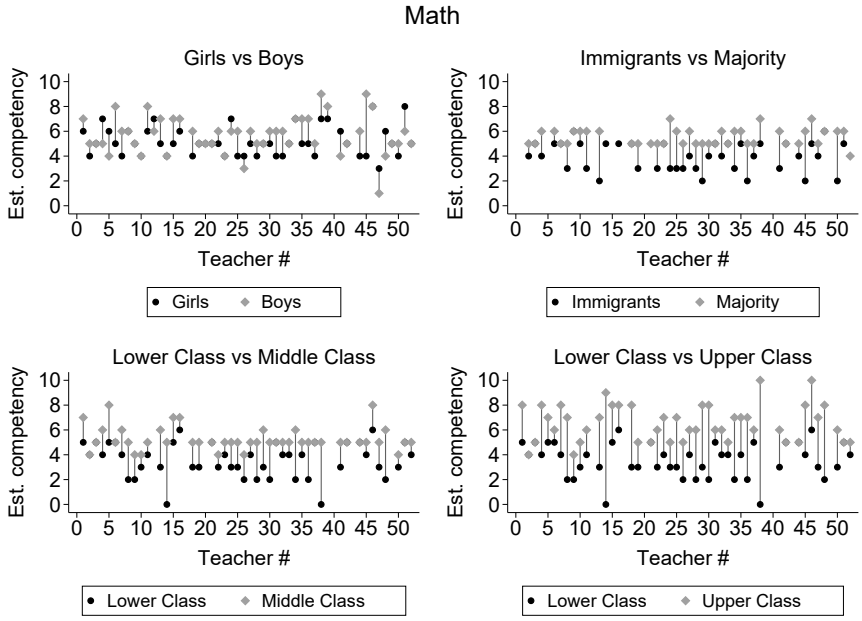


Figure 5.5: Range plots of the differences between teachers' stereotypes of group specific competencies in math by teacher ID.

estimated to have different competencies, some teachers perceive much larger differences between the groups than others, some see no differences at all. Note, however, that these patterns might partly be due to differences in teachers' interpretation of the response scale.

Ranking groups

Teachers not only rank groups correctly on average, the perception of which group of two—if any—is in front, is also mostly correct on the individual level. Actually, for all comparisons shown in figures 5.5 and 5.6—except the one between girls and boys—teachers agree on which group, if any, they expect to be ahead. A notable exception is the difference between girls and boys in math, where some teachers (18.4%) see girls ahead of boys, while about a third of teachers perceives no differences (32.7%) and almost every second teacher believes boys to perform better than girls (49%). Therefore, judged by the empirical studies cited above (Bos, Tarelli, et al., 2012; Bos, Wendt, et al., 2012; Stanat et al., 2012) and the definition of accuracy and bias in section 5.4.2, only about every second teacher holds accurate beliefs about math performance of boys and girls. Over fifty percent hold beliefs that are biased to the disadvantage of

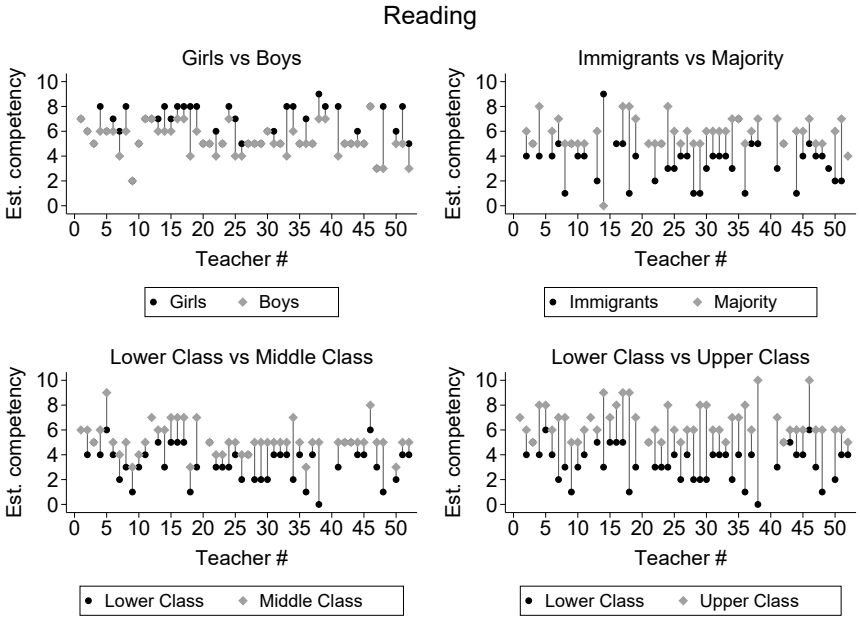


Figure 5.6: Range plots of the differences between teachers' stereotypes of group specific competencies in reading by teacher ID.

boys. Note that those who think that girls outperform boys in math hold least accurate or, put differently, most biased beliefs. In reading, the number for those who hold accurate beliefs is similar: 54 percent of teachers see girls ahead of boys, which is correct (Bos, Tarelli, et al., 2012; Bos, Wendt, et al., 2012; Stanat et al., 2012). The other 46 percent incorrectly believe that there is no difference between the sexes. Thus, these teachers are biased to the disadvantage of girls. Note, however, that, in contrast to the results for math, none of the $N = 52$ teachers believes the reversed order to be true. Thus, the bias to the disadvantage of girls for reading is less severe than the bias to the disadvantage of boys for math.

Most teachers correctly rank students of different social class background: 80% expect students from middle class families to perform better than their lower class peers. In the same vein, 86.7% expect students from upper class families to outperform students with lower class background. 73.3% also see an advantage for upper class students compared to middle class students. Except for the comparison of middle versus upper class, where in both domains one teacher believes that students from middle class families will perform better, the remaining teachers believe that the groups perform equally well. Applying the definition of accuracy and bias from section 5.4.2, these teachers are biased *in favor of* students from the respective lower

class, since all available evidence shows that students from higher social class families perform better on average than students from lower class families (e.g., Bos, Tarelli, et al., 2012; Bos, Wendt, et al., 2012; Stanat et al., 2012). However, judging by the average over all teachers (see section 5.5.1 above), it seems that teachers' stereotypes are biased *to the disadvantage of* students from lower class families and—compared to students from upper class families—to the disadvantage of students from middle class families.

Most teachers correctly expect students from the ethnic majority to outperform immigrants in both reading (85%) and math (76.3%) assessments. Here, too, the remaining percent—except one teacher in case of reading—believes that immigrant and ethnic majority students perform equally well. Thus, the situation is similar to the one of students from different social classes and judging merely from how accurately groups are ranked by the teachers, we might be tempted to conclude that teachers are biased *in favor of* immigrants. However, based on the effect sizes reported in section 5.5.1 above, on average, teachers' stereotypes appear to be biased *to the disadvantage of* immigrants.

Teachers' ranking of students of Turkish and Russian origin are less accurate than the ranking of immigrant and ethnic majority students: For math, only 39.4 percent of teachers expect students of Russian origin to perform better than students of Turkish origin, while a majority of 51.2 percent mistakenly believes that these groups of students will perform equally well and 9.1 percent think that that students of Turkish origin will perform better than those of Russian origin. The numbers are similar for the domain of reading and, thus, similarly biased: only about ever third teacher (31.4%) expects what actually is the case, namely that Russians outperform Turks in reading. Instead, 62.9 percent expect the two groups to perform equally well, 5.7 percent expect Turks to perform better. These results suggest a bias disadvantaging students of Russian origin and, as expected, that teachers perceive these two groups of immigrant students as more similar and homogenous than they really are.

Before I move on to the next section, one last look at yet another visualization of teachers' stereotypes: Figures 5.7 and 5.8 show histograms of all single items. The histograms highlight the variation between teachers. If there were none, each histogram would only show one bar. It is quite clear from figures 5.7 and 5.8 that for both math and reading there are large differences between teachers in the assessment of average competencies of one and the same group of students. In addition to the variation between teachers, the histograms also show that there are differences in the degree to which teachers vary in their assessment of one and the same group. Take immigrants' and majority students' math competencies (figure 5.7), for instance: While teachers'

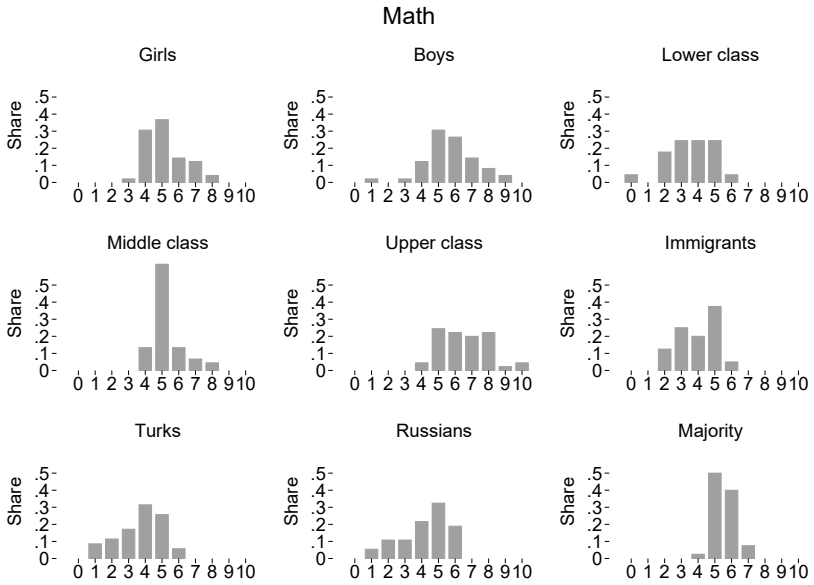


Figure 5.7: Histograms of teachers' stereotypes about group specific competencies in math.

stereotypes of ethnic majority students' competencies are virtually limited to 5, the midpoint of the scale, and 6, immigrants' competencies are estimated to vary considerably.

5.5.3 Item Intercorrelations

While substantively they are certainly less relevant for a study on discrimination in education, item intercorrelations also speak to the validity of an instrument—a valid instrument should show correlation patterns that can be theoretically predicted. Tables 5.1 and 5.2 present item intercorrelations for both domains and all groups. As expected, girls and boys seem to serve as standards of comparison for each other in both domains: The correlations are positive and significant with $r = .42$ for math and $r = .56$ for reading. With regard to social class, things are—also as expected—a little more complex: Interestingly, I observe positive and significant correlations in both math and reading between lower class and middle class (math: $r = .41$, reading: $r = .60$) and between middle class and upper class (math: $r = .60$, reading: $r = .40$) but not between lower class and upper class (math: $r = -.19$, reading: $-.14$).

The correlations among the estimates for immigrants in general and immigrants of Turkish and Russian origin in particular are all statistically significant and range from

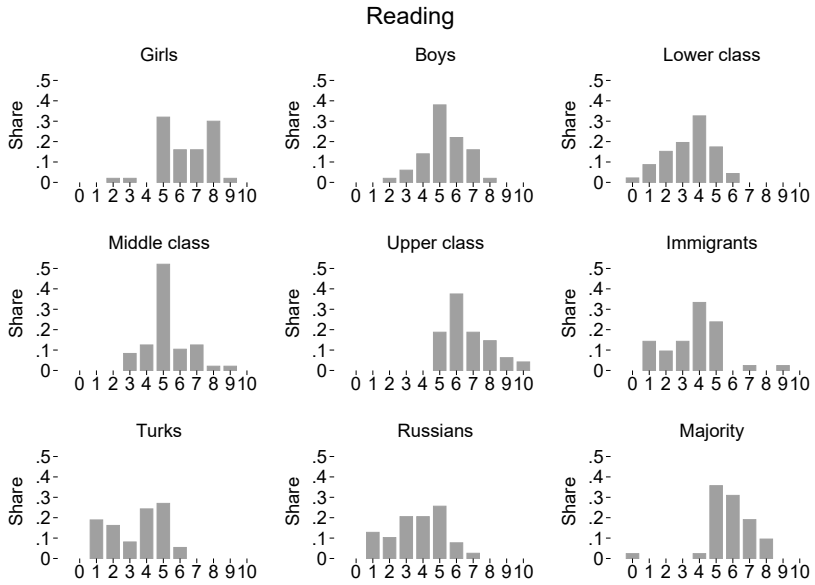


Figure 5.8: Histograms of teachers' stereotypes about group specific competencies in reading.

$r = .56$ to $r = .79$ (see tables 5.1 and 5.2). These strong correlations contrast with low and insignificant correlations between unrelated groups such as girls and boys on the one hand and the different groups of immigrants on the other hand—most of them are virtually zero.

5.6 Summary and Conclusion

In this chapter, I have—building on and extending my joint work with Melanie Olczyk and Georg Lorenz (Wenz et al., 2016)—introduced an item battery to measure teachers' stereotypes about the average competencies in math and reading of different social and ethnic groups, namely girls, boys, students with lower, middle, and upper class background, students of Turkish and Russian origin, as well as students of immigrant origin and majority students. Furthermore, I have raised three research questions, crucial for any study of discrimination in education: First, can we validly measure teachers' stereotypes about different groups of students? If so, do teachers' stereotypes, secondly, differ between groups? And, thirdly, are teachers' stereotypes accurate or more or less biased and, if so, to the disadvantage of which groups of students?

Table 5.1: Item intercorrelations for math.

	Girls	Boys	Lower Class	Middle Class	Upper Class	Immi-grants	Turks	Russians	Majority
Girls	1.00								
Boys	.42**	1.00							
Lower Class	.10	.16	1.00						
Middle Class	.48***	.25	.41**	1.00					
Upper Class	.25	.36*	−.19	.60***	1.00				
Immigrants	.09	−.12	.21	.12	.04	1.00			
Turks	.04	.06	.34*	.06	.03	.59***	1.00		
Russians	.04	.09	.18	.06	.07	.69***	.75***	1.00	
Majority	.57***	.46**	−.12	.28	.34*	.06	.00	.04	1.00

* p <.05, ** p <.01, *** p <.001

Table 5.2: Item intercorrelations for reading.

	Girls	Boys	Lower Class	Middle Class	Upper Class	Immi-grants	Turks	Russians	Majority
Girls	1.00								
Boys	.56***	1.00							
Lower Class	−.03	.36∗	1.00						
Middle Class	.39∗∗	.62***	.60***	1.00					
Upper Class	.61***	.58***	−.14	.40∗∗	1.00				
Immigrants	−.08	.13	.26	.29	.07	1.00			
Turks	−.17	.11	.34∗	.26	.02	.78***	1.00		
Russians	−.06	.00	.29	.19	−.11	.56***	.79***	1.00	
Majority	.33∗	.26	.05	.16	.29	−.28	−.11	.15	1.00

* p <.05, ** p <.01, *** p <.001

Understood as a belief or a set of beliefs about the characteristics, attributes, or behaviors of a particular group or category of people, a stereotype contains more or less accurate beliefs, is held by individuals, and may be measured using implicit or explicit methods. Like many other large-scale assessments in education, the NEPS makes use of paper-pencil self-administered questionnaires, where implicit measures are unfeasible to implement. Therefore, we at pillars 3 and 4 of the NEPS developed an explicit measure of teachers' stereotypes. By means of cognitive interviews we identified a few minor problems respondents might have had with the first version and developed an improved second version of the item battery. This second version was tested in a pilot study with a sample of $N = 52$ second-grade teachers from four German federal states.

I set up quantitative analyses to answer the three related research questions. The analyses show both variation between groups—as a consequence of variation within teachers—and variation between teachers. Both are desirable properties if the instrument is to be used to answer substantive research questions by means of quantitative analyses. Furthermore, the analyses suggest that teachers' stereotypes are quite accurate overall: On average, teachers' stereotypes reflect actual group rankings as judged by numbers reported in the empirical literature. Also, for most group comparisons most teachers correctly rank the different groups. Since teachers are experts with regard to scholastic achievement of different groups of students, I take this as indicative of the instrument's validity. There are two notable exceptions: First, a sizable minority of teachers ranks boys as performing equally or worse as girls in mathematics and, thus, shows a bias against boys. Secondly, students of Russian origin are often seen as performing equally well or even worse compared to students of Turkish origin, while, in fact, the opposite is true.

Evidence based on effect sizes comparing differences in teachers' stereotypes and differences as reported in the literature, suggests that teachers stereotypes are biased to the disadvantage of boys, students from lower social class families or—in comparison with students from upper class families—students from middle class families, immigrants in general, as well as immigrants of Turkish and Russian origin in particular. As expected, these results speak to the general phenomenon of bias in favor of one's—here: teachers'—ingroups, and, hence, to the validity of the instrument. I also find evidence for an outgroup homogeneity effect on the group level: Students of Turkish origin and those of Russian origin are perceived to be more similar than they are according to published studies. What is especially harmful for students of Russian origin—they receive relatively poor assessment in comparison to students of Turkish descent—is yet another piece of evidence for the validity of our measure of

teachers' stereotypes. The fact that estimates for similar or related groups correlate positively, while estimates for unrelated groups do not, also speaks to the validity of the instrument.

Quite obviously, both the instrument and the analysis have their shortcomings and limitations: With regard to the instrument, it cannot be ruled out that teachers adjust their responses towards what they believe to be socially desirable responses. If they do, chances are that they report *smaller* group differences in general and *less negative* stereotypes than they truly hold towards outgroups in particular. However, as shown by cognitive interviews, this should only affect the responses of a minority of teachers. Our question wording seems to successfully hide the true purpose of the instrument and motivate teachers to truthfully report their beliefs. Therefore, the problem of social desirability bias should not be severe.

An important limitation refers to the first research question, namely whether or not teachers' stereotypes can be validly measured by the instrument we developed. The quantitative analyses reported above provide rather indirect evidence that the instrument is indeed a valid measure of teachers' stereotypes. Unfortunately, we could not implement alternative measures of the assessed stereotypes to more directly test the instrument's validity. In this regard future research using the new instrument might provide further insights. However, I think that the quantitative analyses, but especially the theoretical reasoning underlying the item development, and the cognitive interviews provide evidence in favor of the validity of the instrument, since, first and foremost, "validity (as distinct from reliability) is a theoretical concern, not an empirical one" (Lucas, 2008, p. 6). Another limitation is that the instrument does not provide a direct individual-level measure of teachers' perception of the variation within groups. To calculate effect sizes, I assume that the variances between teachers are valid proxies for the average of teachers' perception of within-group variation.

A shortcoming that Wenz et al. (2016) discuss in more detail is the relatively large share of missing values for immigrants, Turks, Russians, and majority students (see figure 5.4 for the numbers). However, it might be that teachers who did not answer these items have less or no experience with students of such origin. If so, the larger share of missing values for these groups would be less problematic, since teachers' stereotypes should affect only the outcomes of students they actually teach. This relates to the problem already mentioned in chapter 4 that teachers who contribute to the bias against particular groups of students in the sample as a whole may not teach students from the respective group. However, I did not restrict the sample of $N = 52$ any further to restrain keep the sampling error low and statistical power as high as possible. Whether teachers who did not report a particular stereotype actually have

less or no experience with students from the group in question and whether teachers who show biases against particular groups of students do in fact teach students from the respective groups, could and should be tested in future research using the data from the scientific use files from the NEPS that feature larger sample sizes and were published after the analyses for this chapter were conducted.

Despite the shortcomings and limitations that the new instrument and the analyses presented in this chapter certainly have, I am confident that the new instrument—the first explicit measure of teachers' stereotypes in a panel study on education—is a valid measure of teachers' beliefs about the average competencies of different groups of students. I also think that the analyses in this chapter provide sound evidence for the conclusion that, while they are not grotesquely off, teachers' stereotypes are somewhat biased to the disadvantage of boys compared to girls, and more so to the disadvantage of students from lower social class families, immigrants in general, as well as immigrants of Turkish and Russian origin in particular—always to the respective comparison group of students from higher social classes or ethnic majority students. Applying these stereotypes to individual students in situations such as grading or recommending tracks at the end of elementary school should, thus, not only lead to individual discrimination, but also group discrimination and, hence, help explain inequality in German education.

6 Discrimination in German Education: An Experiment[34]

> Establishing that [...] discrimination did or did not occur requires causal inference.
>
> *(Blank et al., 2004)*

> For obtaining causal inferences that are objective, and therefore have the best chance of revealing scientific truths, carefully designed and executed randomized experiments are generally considered to be the gold standard.
>
> *(D. B. Rubin, 2008)*

In this chapter I address the question of whether there is discrimination in German education along the lines of ethnicity, social class, and gender by means of an experimental study. Based on the conceptualization proposed in chapter 2, I am interested in discrimination as causal effect of an information about or signal sent out by a student on how this student is treated by a teacher. Here, I am interested in the causal effects of signals that carry information about a student's ethnicity, social class, and gender. With regard to ethnicity, I focus on students of Turkish origin in comparison to students of the German ethnic majority. Regarding social class, I am interested in the treatment effects for signals of lower class versus upper middle class backgrounds. As for gender, I look at the contrast between girls and boys.

6.1 Observational Studies

As briefly mentioned in chapter 1, there are not many studies that explicitly investigate ethnic discrimination or discrimination against immigrants in German education and even less that discuss sex or gender discrimination or discrimination by virtue of a students socioeconomic background. In her seminal study on ethnic discrimination in

34 This chapter is based on joint work with Kerstin Hoenig (Wenz & Hoenig, 2020). While there are only minor differences between this chapter and Wenz and Hoenig (2020), I recommend reading and citing Wenz and Hoenig (2020) instead of this chapter.

German education, Kristen (2006b) does not find evidence for ethnic discrimination to the disadvantage of students of Turkish or Italian background in both grades and track recommendations at the end of elementary school. Only one particular model specification yields a statistical significant disadvantage for students of Turkish backgrounds in German grades after controlling for test scores (Kristen, 2006b, p. 90, footnote 7). Evidence from other studies is more mixed: Conditional on relevant controls—usually including a measure of socioeconomic background—, other studies find disadvantages in terms of statistically significant negative ethnic residuals in grades or track recommendations (e.g., Gresch, 2012; Kiss, 2013; Lüdemann & Schwerdt, 2013), but some even find positive residuals (e.g, Gresch, 2012). Virtually all observational studies on ethnic discrimination in education also find non-significant ethnic residuals in some models, depending on the exact specification. For a review on ethnic discrimination in German education see Diehl and Fick (2016).

The evidence from observational studies with regard to social class is much stronger: Even though studies typically do not investigate social class discrimination explicitly (cf. T. Schneider, 2011)—which is, given the large socioeconomic disparities in educational achievement rather surprising in and of itself—there are numerous studies whose findings can be interpreted as evidence for discrimination in grading and, even more so, track recommendations by virtue of students' social class background: Conditional on competencies and other relevant covariates, these studies find that teachers recommend or prefer lower tracks for students from lower class families (e.g., Arnold et al., 2007; Bos, Tarelli, et al., 2012; Bos, Wendt, et al., 2012; Ditton, 2013; Ditton et al., 2005; Maaz et al., 2011; Maaz et al., 2010; T. Schneider, 2011; Wendt et al., 2016).

With regard to gender, some studies find that boys receive lower grades conditional on test scores and other relevant controls (e.g., Hochweber, 2010; Maaz et al., 2011), other studies do not find such an effect (e.g., Wendt et al., 2016). By and large, observational studies suggest that, if anything, discriminatory grading to the disadvantage of boys is rather small in effect size. Similarly, some studies find statistical significant disadvantages of boys remaining in teachers' track recommendations or track preferences (e.g., Arnold et al., 2007; Ditton et al., 2005), but others—typically more recent studies—find no such effect (e.g., Bos, Tarelli, et al., 2012; T. Schneider, 2011).

6.1.1 Limitations of Observational Studies

Residual estimates of discrimination Oaxaca (1973) hinge on the assumption that all relevant controls have been included in the model and measured without error. An

important case in point is the students' actual performance at school, be it in the classroom, homework assignments, or in tests and quizzes. Results from standardized competence tests usually serve as a proxy, but these do not perfectly represent the students' performance in class and that is the basis of teachers' evaluations. Note, that such residual estimates might either over- or underestimate discrimination due to under- or overcontrolling of key variables, respectively (Holzer & Ludwig, 2003). With regard to discrimination in education, an obvious example for undercontrolling and, thus, overestimation of discrimination, is the lack of valid and reliable measures of classroom participation. Overcontrolling might happen in the face of racially biased test scores: If, for example, ethnic minority students' test scores are negatively affected by the students' fear of confirming negative stereotypes about their intellectual means, that is controlling for these racially biased test scores would lead to an underestimation of discrimination (Croizet, 2008; Steele & Aronson, 1995).

6.2 Experimental Studies

One proposed solution to the problems of observational studies is an experimental research design, in which randomization, if successful, prevents both unobserved heterogeneity and self-selection of any kind. In experiments conducted in the lab or in the field, actual performance is under the control of the researcher. Natural experiments typically exploit some kind of "natural" randomization.

6.2.1 International Studies

Most research has been conducted in the US, where a tradition of experimental research on discrimination in education dates back to the 1970s (e.g., DeMeis & Turner, 1978; Feldman & Orchowsky, 1979; Harari & McDavid, 1973; Rubovitz & Maehr, 1973; Taylor, 1979). More recent contributions to US literature have focused on using data from larger field experiments, natural experiments, or similar quasi-experimental designs (e.g., Dee, 2004b; Figlio, 2005). But there are also experimental studies from Sweden (Hinnerich et al., 2011), the Netherlands (van Ewijk, 2011), Israel (Lavy, 2008), India (Hanna & Linden, 2012), and Germany (e.g., Schulze & Schiener, 2011; Sprietsma, 2013). They typically employ a design similar to the one proposed in the present study: participants, usually teachers, but quite frequently also students, are asked to evaluate the performance, such as an essay, a written exam or an audio recording, of a subject whose characteristics, e.g., gender, social class, ethnicity or immigrant background, are varied randomly. Often, this is done by varying the subject's name, but some stud-

ies provide pictures or extensive vignettes (e.g., Hanna & Linden, 2012; Schulze & Schiener, 2011). Another popular type of design are natural experiments. Lavy (2008), for instance, compares blind and non-blind examination data from official registers. Hinnerich et al. (2014) explicitly hired teachers to blindly grade exams to compare this blind score to the non-blind grade as given by the students' teachers.

The majority of these studies finds evidence for discrimination against under-privileged groups on various dimensions, including race (Dee, 2004b; DeMeis & Turner, 1978; Feldman & Orchowsky, 1979; Rubovitz & Maehr, 1973), ethnicity (Sprietsma, 2013; van Ewijk, 2011), immigrant background (Hinnerich et al., 2014), caste (Hanna & Linden, 2012), attractiveness (DeMeis & Turner, 1978; Harari & McDavid, 1973), and gender (Lavy, 2008; Lindahl, 2016). However, Lindahl (2016) finds the opposite effect when it comes to migration background: students with a non-native name are favored by teachers when it comes to deciding who gets a school leaving certificate. Van Ewijk (2011), whose design is very similar to the present study, finds no discrimination at all in essay grading, but lower expectations for immigrants' future academic success.

6.2.2 Evidence From Germany: The Study by Sprietsma (2013)

As a replication and extension of van Ewijk (2011), Sprietsma (2013) assesses teachers' biases in grading and recommendations for secondary school tracks. To this end, Sprietsma (2013) randomly assigns names that signal a Turkish immigrant background and names that signal German heritage to four sets of essays that were sent out to 3500 schools in two otherwise unspecified regions in Germany. $N = 88$ teachers sent back the graded essays and filled in questionnaires. Sprietsma (2013) finds, based on linear regression models, a statistically significant bias in grading of about .1 standard deviation to the disadvantage of what appeared to be Turkish students to the teachers. Using so-called feeling thermometers, she does not find—in contrast to what I find in chapter 4 using measures of social distance—statistically significant negative prejudice against Turks except for the group of teachers that reports to have little experience in teaching students of immigrant background.

However, Sprietsma (2013) finds a discriminatory bias in grades assigned to essays of about $.1$ standard deviations to the disadvantage of those essays that were supposedly written by students with a Turkish immigrant background. She also finds that teachers are on average 11 percentage points less likely to recommend *Gymnasium* to a student with a Turkish name compared to a student with a German name but no effect for *Realschule*. While the results from this first larger experimental study in Germany on ethnic discrimination by teachers are certainly informative with regard

to the question of whether or not ethnic discrimination exists in German education, it has—at least—two noteworthy limitations: A first limitation that I will discuss in more detail below is that, by comparing average Turkish names to average German names, Sprietsma (2013) cannot distinguish between ethnic discrimination and social class discrimination. Secondly, the sample of $N = 88$ teachers was recruited out of 3500 schools that were sampled and, thus, "the response rate was extremely low and [the] sample is not representative of the primary school teacher population of these regions" (Sprietsma, 2013, p. 529).

6.2.3 Problems of Experimental Studies

While the experimental design is often described as the gold standard of causal analysis (see, e.g., Gangl, 2010; Imbens & Rubin, 2015; Morgan & Winship, 2015; D. B. Rubin, 2008), it is not without problems. Some of theses problems are design-immanent, some are mainly due to the way researchers handle experimental designs. The following three problems strike me as the most severe problems of experiments on discrimination.

Sampling and sample of analysis

There are two distinct problems most experimental studies suffer from that concern the sample these studies rely on. I have already mentioned both in chapter 4 and chapter 5. First, many experimental studies in the field of discrimination rely on convenience samples, often drawn from populations such as university students in education programs or preservice teachers (e.g., Bonefeld & Dickhäuser, 2018; Glock et al., 2015; Schulze & Schiener, 2011). In their sample, Glock et al. (2015) even mix teachers from one European country with preservice teachers from another. Therefore, these studies have rather low external validity and, in fact, can be shown—as I have done in section 4.4.1—to result in rather biased estimates of population parameters.

Secondly, a point I have also made in section 2.4.1 but have not seen addressed in any study is based on insights by Becker (1957/1971), nicely summarized by Heckman (1998) in the following sentence: "finding a discriminatory effect of race or gender at a randomly selected firm does not provide an accurate measure of the discrimination that takes place in the market as a whole" (Heckman, 1998, p. 102). In the labor market, the difference is mainly due to self-selection of employees into non-discriminating firms; a behavior that causes segregation (Becker, 1957/1971). In elementary education in Germany, differences in the level of discrimination between

an average or typical teacher and a teacher that actually teaches students of a particular background—e.g., students of Turkish background—might more likely arise from teachers self-selecting into schools with different shares of students of such background or from a change in attitudes and beliefs as a consequence of teaching such students. Self-selection of students into elementary schools is heavily restricted in most states in Germany and, hence, should play less of a role. I have not seen this exact point being explicitly addressed in any empirical study on discrimination in education, although van Ewijk (2011) makes a similar point and oversamples schools with a share of at least 25% non-ethnic Dutch students.

Effect heterogeneity across the competence distribution

From models of statistical discrimination theory we know that discrimination may differ along the distribution of observed performance, y (Aigner & Cain, 1977; Phelps, 1972). In fact, if teachers engage in reliability-based statistical discrimination it is possible that not only does the discriminatory effect vary but also change its direction over the range of y. Therefore, studies that only assess discrimination at one point along the distribution of y are severely limited with regard to what can be inferred from them about discrimination in the market or sector of interest. Put differently, without making additional assumptions, these studies cannot say anything about the average level or direction of individual discrimination nor can they say anything about group discrimination—not even which group is on average suffering from it (also see Heckman, 1998; Heckman & Siegelman, 1993; Neumark, 2012, for this argument).

Confounding ethnic and social class discrimination

That ethnic or racial discrimination might at least partly be a problem of social class discrimination is an insight that dates back to the earliest days of research in race relations (e.g., Myrdal, 1944, p. 75). Blalock (1967), for example, devotes an appendix to the questions whether racial prejudice is essentially class prejudice or to what degree race and class attitudes are interchangeable (Blalock, 1967, pp. 199–203). He summarizes the problem as follows:

> The problem of distinguishing between racial prejudice and class attitudes arises because of the fact that ethnic and racial backgrounds are among the important criteria used to determine one's general status. *As long as minority membership remains among the defining criteria of class position it*

> *will indeed be difficult to separate the two phenomena empirically*. (Blalock, 1967, pp. 201-202, his emphasis)

Put differently: Because ethnicity and social class are confounded, measures of ethnic prejudice or discrimination might be confounded with social class prejudice or discrimination. And, indeed: What was true back then is still true today in virtually all societies. In Germany, for example, immigrants with Turkish background are overrepresented in the lower classes and have generally worse labor market outcomes than the ethnic majority (Below, 2007; Büchel & Frick, 2004; Kalter, 2008; Kogan, 2004, 2007).

In conjunction with the standard research design of manipulating ethnicity by varying names or pictures (e.g., Bertrand & Mullainathan, 2004; Deming et al., 2016; Glock & Klapproth, 2017; Jacquemet & Yannelis, 2012; Sprietsma, 2013; Weichselbaumer, 2016), this, as I argue, may pose a serious problem. The problem lies in the selection of stimuli and the corresponding signal: By selecting a name or picture typical for an average minority group member and an average majority group member, the stimuli signal not only ethnicity but also all societal correlates of which, in case of ethnicity, the most important is arguably social class. So, in contrast to most studies relying on observational data, social class is not held constant in the experimental manipulation. Thus, it is possible that any discrimination found in these studies is—at least partly—the result of class differences, not ethnicity (for this and similar arguments see Bertrand & Mullainathan, 2004; Figlio, 2005; Fryer Jr & Levitt, 2004; Gaddis, 2015, 2017a, 2017b). Figure 6.1 visualizes the problem: Names, N, are determined by both ethnicity at conception, E_C, and social class at conception, C_C, and affect, that is, send, both ethnic signals, E_S, and social class signals, C_S.

To my knowledge, the present study was the first experimental study that made a conscious effort to disentangle class and ethnic discrimination in education using names as stimuli. Tobisch and Dresel (2017)—who cite Hoenig and Wenz (2013) in a correction (Tobisch & Dresel, 2020)—have since published a study with a similar design that replicates the main findings of our study as reported in Hoenig and Wenz (2013) and in this chapter below.

Note that while vignette designs that make use of extensive descriptions of students (Glock et al., 2015; Hanna & Linden, 2012; Schulze & Schiener, 2011) may help to address the problem, they have other limitations of which the most important is that they create rather artificial settings that make it difficult to provide a reasonable cover story that hides the true purpose of the study. I have not reviewed the vignette study by Schulze and Schiener (2011) in detail above as it relies on a student sample. Its finding that immigrant background—operationalized via language spoken at home—

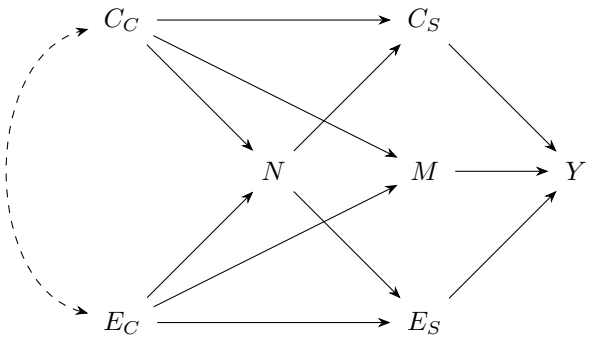


Figure 6.1: Stylized directed acyclic graph (DAG) showing the problems of identifying ethnic discrimination, $E_S \rightarrow Y$, and social class discrimination, $C_S \rightarrow Y$, using names, N, as treatments. A randomized assignment of names, N, blocks the backdoor paths through ethnic and social class at conception, E_C and C_C. However, if N carries both ethnic and social class signals, E_S and C_S, the backdoor paths $E_S \leftarrow N \rightarrow CS \rightarrow Y$ and $C_S \leftarrow N \rightarrow E_S \rightarrow Y$ remain open.

does not affect the probability of a recommendation for *Gymnasium* independent of the parents' education, is nevertheless an interesting finding on the backdrop of the discussion about the confounding of ethnicity and social class.

6.3 The Situation at the End of Elementary School in Germany

Discrimination by teachers may occur in at least three situations: First, day-to-day classroom interactions, for example by calling more on some students than others. Secondly, in the evaluation of a specific performance, for example when grading a test or assignment. And, thirdly, in the evaluation of a student's general potential or in the formation of expectations about a student's development—such expectations should play a major role for all kinds of treatments including decisions about ability grouping or tracking and may lead to self-fulfilling prophecies. In this experiment I take a look at the two latter types—grading and expectations. I rely on the theories I discussed in chapter 3 and the findings regarding teachers' prejudices and stereotypes in chapters 4 and 5 to derive hypotheses depending on the situation. Each of the two situations features a different logic with different conditions that allow to indirectly test the mechanisms suggested by different theories.

In chapter 1, I have argued and cited evidence for the importance of the first transition in German education, namely the transition from elementary school to secondary school. This transition provides us with a test case scenario in which both teachers'

grading and expectation formation matter a lot. Even though the 16 German federal states are responsible for their education policy, their education systems are actually fairly similar (Helbig & Nikolai, 2015). All children start elementary school around the age of six. Usually, a single teacher teaches the main subjects and there is no formal ability grouping or streaming. In most states, students are tracked into different school types after four years of elementary school when they are on average 10 years old.[35] The number and specifics of tracks differ between states, but in all states, the highest track is the *Gymnasium*, which leads to the *Abitur*, the highest secondary degree and entrance ticket to university. In all states, elementary school teachers give official recommendations suggesting the track they believe would be ideal considering the child's potential. The major—or, in some states, only—determinants of these recommendations are grades, but teachers are, legally or empirically, asked to consider the child's overall potential. These track recommendations are legally binding in some federal states, but can be overruled by parents in others (Helbig & Nikolai, 2015).

Because tracking between different school types occurs unusually early in the German education system and because a student's track is largely determined by the teacher's recommendation, discrimination by elementary school teachers can have especially severe consequences for children's educational attainment in Germany. However, my theoretical and methodological contributions are of much broader significance, as both the grading of students' performance in a non-anonymous setting and some form of tracking, streaming, or ability grouping, dependent largely on teachers' evaluations and grading, takes place in virtually all education systems. The only unusual feature of the German system is the young age at which between-school tracking occurs (OECD, 2010). Furthermore, and as discussed in chapter 5, teachers' expectations may turn into self-fulfilling prophecies that can be especially harmful to students from stigmatized groups (Jussim, 1989; Jussim et al., 1996; Jussim & Harber, 2005; Jussim, Robustelli, et al., 2009; Lorenz et al., 2016).

6.4 Hypotheses

The most obvious and, as I am going to show in this section, theoretically important difference between forming expectations and grading a manifest performance such as a written essay is the amount and reliability of individuating information available to the teacher. While in principle an essay provides the teacher with all information

35 In two states—Berlin and Brandenburg—children are tracked after six years at the age of 12.

needed to grade it, tracking or grouping decisions are based on expectations that are themselves necessarily based on imperfect information about the latent construct of ability or potential and—especially in situations of explicit between school tracking—a yet unobserved future. As I intend to provide indirect evidence on the mechanisms governing teachers' judgments, in this section I briefly recapitulate theoretical mechanisms that could govern discrimination in grading and expectation formation as discussed mainly in chapter 3 but also in chapters 4 and 5. I then deduce hypotheses from different theories and perspectives.

6.4.1 Tastes, Prejudice, and In-Group Favoritism

As discussed in chapter 3, Becker (1957/1971)'s theory of *taste discrimination* and *social identity theory* (SIT; Tajfel, 1982; Tajfel & Turner, 1986) are similar in the sense that both do not put emphasis on the role information or situational ambiguity plays but more on intergroup relations and common group membership. Becker (1957/1971) suggests that discrimination in favor of or to the disadvantage of a member of a particular group compared to a member of another group occurs whenever an actor has different discrimination coefficients (DCs) for the two corresponding groups. SIT suggests that both personal and social identity determine a person's self-esteem and self-concept. Since humans strive to maintain a positive self-concept and enhance self-esteem, they also strive for strive for positive personal and social identities. The social identity might be positively affected by favoring ingroups and ingroup members or by derogating outgroups and outgroup members. This mechanism not only explains discrimination but also why people tend to hold negative prejudices towards outgroups and outgroup members.

In chapter 4 I have reviewed studies and provided own evidence that suggests that German teachers indeed hold negative prejudices against Turks in general and students with a Turkish background in particular. The social distance measures I used as global measures of prejudice can also be understood as measures of Becker (1957/1971)'s DC. Since elementary school teachers in Germany are also overwhelmingly female instead of male and are themselves—as academics employed in the public sector—members of the service class, boys and children from lower social class backgrounds are outgroup members that teachers arguably also hold negative prejudices against. In sum, both SIT and Becker (1957/1971) expect discrimination to occur largely regardless of the amount and reliability of individuating information against outgroups and outgroup members.

However, before I deduce hypotheses from these perspectives, please recall that in

chapter 4 I have also suggested that the mechanism, Becker (1957/1971) proposes to explain discrimination with regard to wage setting or hiring decision—namely "disutility caused by contact with some individuals" (Becker, 1957/1971, p. 15)—, cannot be straightforwardly applied to all kinds of situations in education including essay grading and recommending tracks. While I stick to this interpretation, I would like to remind the reader, that in the economic and sociological literature Becker (1957/1971)'s model is often interpreted more generally as a model in which tastes, prejudice, or ingroup favoritism govern human behavior (e.g., Hanna & Linden, 2012; Kristen, 2006b; Sprietsma, 2013; van Ewijk, 2011). From this perspective, the predictions of Becker (1957/1971)'s theory of taste discrimination and social identity theory coincide. But, as as a theory test, I suggest that the following hypotheses are more informative with regard to SIT than Becker (1957/1971):

Hypothesis 1_a*:* German teachers discriminate in *both* essay grading and expectation formation by virtue of students' ethnic background to the disadvantage of students with a name signaling a Turkish background.

Hypothesis 1_b*:* German teachers discriminate in *both* essay grading and expectation formation by virtue of students' social class background to the disadvantage of students with a name signaling a lower social class background.

Hypothesis 1_c*:* German teachers discriminate in *both* essay grading and expectation formation by virtue of students' gender to the disadvantage of students with a name signaling male gender.

6.4.2 The Role of Imperfect Information and Ambiguity

Empirical evidence and more recent theoretical contributions point to situational moderators of the link between categorization and the application of stereotypes or prejudice and, hence, discrimination. I discussed three rather different theories that fall in this camp: Statistical discrimination theory (Aigner & Cain, 1977; Arrow, 1973; Phelps, 1972), the continuum model (Fiske et al., 1999; Fiske & Neuberg, 1990), and the theory of aversive racism (Dovidio & Gaertner, 2008; Gaertner & Dovidio, 1986). However, mechanisms and theoretical reasoning differ considerably among these approaches as I have already discussed in detail in chapter 3.

Statistical discrimination (e.g., Aigner & Cain, 1977), for instance, points to imperfect knowledge as the key reason for why rational decision makers discriminate on the basis of observable group characteristics. Following statistical discrimination theory, teachers should be expected to construct a weighted average of observed individual performance and assumed group ability to estimate a student's individual ability: The

lower the reliability of the individual information, the further the estimate is pulled towards the assumed group mean, that is, towards the teacher's stereotype.

The continuum model (Fiske et al., 1999; Fiske & Neuberg, 1990) acknowledges available information as one of "two primary factors" in more or less category or stereotype driven and, hence, more or less discriminatory judgments and behavior. If the target is of minimal interest or relevance for the perceiver in the very moment of categorization, perceivers are motivated to allocate attention to individuating information and move down the continuum from category-based "affect, cognitions, and behavioral tendencies" toward a "piecemeal integration" of individual attributes (Fiske et al., 1999, p. 233). Of course, this process of recategorization and, eventually, piecemeal integration may only be started if the available information is rich enough and the perceiver has the time and the cognitive capacity to take it into account.

The theory of aversive aversive racism (Dovidio & Gaertner, 2008; Gaertner & Dovidio, 1986) suggests that in modern societies, where negative prejudice and discrimination to the disadvantage of ethnic and social minorities is condemned by a majority of people, many people are motivated to uphold a positive self-image as unprejudiced nondiscriminator but at the same time hold negative implicit prejudices and stereotypes about outgroups and outgroup members. These aversive racists are expected to discriminate only in situations that are ambiguous enough to not reveal the discriminatory behavior to the aversive racists themselves and others.

As for teachers' stereotypes and attitudes, I have reviewed evidence provided by others and provided own evidence in chapter 5 that teachers in Germany hold stereotypes that are biased to the disadvantage of Turkish students, students from lower social class families, and boys. Therefore, in case teachers rely on a decision algorithm as proposed by statistical discrimination theory and individual information is not perfectly reliable, we should expect the special case of "error discrimination" (England & Lewin, 1989). Biased stereotypes as well as implicit and explicit negative prejudices (see chapters 5 and 4) are relevant and may lead to discrimination if teachers rely on a category-based judgment instead of a piecemeal integration of individual attributes (Fiske et al., 1999) or can get away with a discriminatory response (Gaertner & Dovidio, 1986).

As for the logic of the situation teachers find themselves in when grading a written essay, sufficiently motivated teachers should *not* show any discriminatory biases, since all relevant information is available (Aigner & Cain, 1977; Fiske et al., 1999) and a discriminatory bias is hard to hide (Gaertner & Dovidio, 1986).

However, when the same teachers need to predict future development of students when recommending tracks, the available information based on an essay might not

be perceived as perfectly reliable (Aigner & Cain, 1977), or—put differently—might not be rich, diagnostic, or clear enough (Dovidio & Gaertner, 2008; Fiske et al., 1999; Fiske & Neuberg, 1990; Gaertner & Dovidio, 1986). In this case, teachers should make use of beliefs about group means and stereotypes in general (Aigner & Cain, 1977), may have not enough information to go all the way from a category-based response to a piecemeal-based response (Fiske et al., 1999; Fiske & Neuberg, 1990), and may take the opportunity to hide a judgment based on stereotypes or prejudices behind vague information and the ambiguity of the situation (Gaertner & Dovidio, 1986).

Based on how statistical discrimination theory, the continuum model, and the theory of aversive racism acknowledge the reliability of information and the ambiguity of situations as moderating factors that increase the likelihood of category based judgments, I deduce the following hypotheses:

Hypothesis 2_a*:* German teachers *do not* discriminate when grading a written essay *but do so* when forming expectations by virtue of students' ethnic background to the disadvantage of students with a name signaling a Turkish background.

Hypothesis 2_b*:* German teachers *do not* discriminate when grading a written essay *but do so* when forming expectations by virtue of students' social class background to the disadvantage of students with a name signaling a lower social class background.

Hypothesis 2_c*:* German teachers *do not* discriminate when grading a written essay *but do so* when forming expectations by virtue of students' gender to the disadvantage of students with a name signaling male gender.

6.4.3 Further Thoughts on What to Expect

Instead of deriving more concrete hypothesis, in this section I offer some further thoughts on what different theoretical models might predict for the different situations with a special focus on effects at different points in the distribution of observed performance.

Clearest guidance in this regard comes from statistical discrimination theory. As discussed elsewhere in this dissertation (see, e.g., section 3.1.2 again), statistical discrimination models suggest that discrimination may differ along the distribution of observed performance (Aigner & Cain, 1977; Phelps, 1972). If teachers attach different reliabilities to performance signals from different groups—e.g., students of Turkish background and those without immigrant background—, individual discrimination should vary over the range of observed performance. However, without knowing the group-specific reliabilities, where exactly teachers see the performance presented in the experiment—i.e., in which part of the distribution—, and how risk-averse teach-

ers really are (cf. Maaz et al., 2008), it is difficult to derive concrete hypotheses. This is why I will not do so but simply remind the reader that and why interaction effects with essay quality will be specified and are of great interest.

With regard to the other two theories that highlight situational moderators such as the richness of information and ambiguity, it might certainly be the case that essays of different quality relate to these mechanisms.

The continuum model (Fiske et al., 1999; Fiske & Neuberg, 1990) proposes that category-based affect, cognition, and behavioral responses are the default mode of human cognition. Only if the initial categorization of a person does not seem to fit the data—here: the written essay—, a recategorization process is started that potentially leads all the way down to a piecemeal integration of the available data and, thus, to an individuating response. It could be, for example, that the better of two essays is so good, that the teacher finds it difficult to achieve a fit to the stereotype of a student with Turkish background and, instead of a category-based response, looks very careful for individuating information and behaves accordingly. At the same time, the teacher might achieve a good fit for this essay and the stereotype of a German student from an upper middle class family. This might result in similar predictions for these students and no discrimination on the basis of ethnic signals that could disadvantage the Turkish student. If at the same time, the bad essay's nature is such that it allows the teacher to proceed with a category-based response in both cases—because the essay is rather average, not very good, not very bad—there should be discrimination on the basis of the ethnic signal and corresponding stereotypes. Of course, this example also works the other way around—with a very bad and an average essay. In either case, the resulting pattern of this scenario would be an interaction effect of discriminatory responses with essay quality.

Similarly, relying on mechanisms from aversive racism theory we might also predict such an interaction effect. If performance is clearly very good or very bad, aversive racists will not apply stereotypes and prejudices in their judgments and behavior.

6.5 Experimental Design

I designed an online experiment to identify and estimate ethnic discrimination, social class discrimination, and gender discrimination in grading of a specific assignment and in teachers' expectations. In 2010 I collected the data together with Kerstin Hoenig and Anne Landhäußer. Test subjects were elementary school teachers from the German federal state of Baden-Württemberg who taught German at the time of data collection. I limited the sample to teachers from one federal state because cur-

ricula and grading standards differ between states. I employed a $2 \times 2 \times 3$ factorial design, varying essay quality, the student's gender, and the student's social and ethnic background. Gender and background were varied by random assignment of names to the essays and participants. In addition to grading one essay each, teachers were asked to answer a short questionnaire.

6.5.1 Sampling and Contact

The main goal in sampling was to increase external validity and reduce bias in my estimates compared to most previous research. To this end, I drew a random sample of 720 schools from all primary schools in the state of Baden-Württemberg, both public and private, and contacted them via e-mail. The recipient—in most cases, probably either the school's principal or secretary—was asked to forward the e-mail to all teachers at the school who taught German at the time. As an incentive to participate, I put up a lottery of three gift certificates for an online book store worth EUR 20.00 each, as well as the option to receive information about results. $N = 237$ teachers participated in the survey.[36]

6.5.2 Essays

Each teacher was presented with one of two essays of different quality. By varying essay quality, I address the shortcoming of some previous studies that assess discrimination merely at one point in the distribution of y. Essays were obtained from a fourth grade class from Baden-Württemberg[37]. They were about 200 words long and were based on the assignment to write a story around a given title. The two essays for our experiment were chosen based on the results of a pretest in Bavaria, a state whose education system and educational outcomes are similar to that of Baden-Württemberg. The pretest also served as a test of the sampling and contacting procedure of the main study. 27 teachers from randomly sampled Bavarian elementary schools took part in the pretest. They graded each essay without receiving any information about the supposed author except for age and grade level. Additionally, they were asked to guess the child's gender and to answer a short questionnaire about their own teaching expe-

36 Due to restrictions by the state's Ministry of Education, which had to approve the study, I was not allowed to ask teachers the name of their school in the questionnaire. This unfortunately means that I am unable to account for potential clustering of teachers by schools in my analyses.

37 I would like to thank Anne Landhäußer for collecting the essays.

rience. Based on the results of the pretest, I selected two essays that were of different quality and comparatively gender-neutral, as I would assign both male and female author's names to each essay in the main study.

6.5.3 Names

To identify and estimate ethnic and class discrimination, I chose one male and one female name each that signal a German upper middle class (Jakob, Sophie), German lower class (Justin, Jacqueline), or Turkish background (Ayse, Murat), respectively. I made sure that the German names I selected are about equally prevalent in the birth cohort of 2000 and that none of the names are linked to a certain geographical region in Germany. Although there are apparently typical upper and lower class Turkish names, I have reason to believe that German teachers are simply not familiar enough with Turkish culture to recognize the difference. Therefore, I cannot vary class background independently of ethnic background by name manipulation. Instead, I assume that a Turkish name indicates a class background that is comparable to that of German lower class names. This assumption was tested in a manipulation check on a different sample (see below). With regard to the definition of ethnic discrimination as causal effect of an ethnic information or an ethnic signal, the two causal states whose difference define this causal effect are names that signal Turkish background and names that signal German lower class background. The differences between how these groups are treated by teachers are interpreted as evidence of ethnic discrimination. The differences between names signaling German upper middle class and German lower class are interpreted as social class discrimination. I return to the question whether these definitions are the only meaningful and defendable ways of defining ethnic discrimination and social class discrimination in such an experimental design.

The alternative to using names as treatments would have been a design based on comprehensive vignettes that explicitly include child background characteristics. I decided against the use of vignettes because these create a rather artificial setting in which teachers are bound to ask themselves why the researcher provided them with this information in this form—vignettes typically use extensive descriptions of a situation or person. In contrast, names make for a much more subliminal stimulus, that—embedded in a reasonable cover story—should reduce social desirability bias to a minimum. Also, among other more substantive reasons, I motivated the present study by pointing to methodological problems of research designs that use names as stimuli without addressing the question of whether or not ethnic or racial signals may be confounded with social class signals.

Manipulation check

The names underwent a manipulation check using a separate sample of elementary school teachers ($N = 75$), who were asked to guess the migration and class background (upper, middle, or lower class) of each name. As intended, the vast majority of teachers indicates that Murat (69%) and Ayse (70%) have a Turkish background, whereas Jakob (94%), Sophie (97%), Justin (99%), and Jacqueline (92%) are virtually unanimously identified as German.[38] Remarkably, a sizable minority identifies Murat (23%) and Ayse (24%) as German.

Regarding social class, Sophie and Jakob are believed to come from families with an upper class (Sophie: 53%, Jakob: 59%) or middle class (Sophie: 40%, Jakob: 36%) background. As expected, for Jacqueline and Justin the pattern is reversed: these names are perceived predominantly as names held by children with a lower social class background (Jacqueline: 71%, Justin: 67%). However, a sizable minority of teachers categorizes them as middle class (Jacqueline: 21%, Justin: 25%) or even upper class (both 8%). The Turkish names are also perceived as being most likely to be names from students with a lower social class background (Ayse: 52%, Murat: 53%), followed by middle class (Ayse: 39%, Murat: 36%) and upper class (Ayse: 9%, Murat: 11%). Yet, fewer teachers report to perceive Ayse and Murat as lower class than Jacqueline and Justin.

Overall, teachers tend to perceive the selected names as intended. However, there are potential problems: If a sizable minority of participating teachers really perceive students with Turkish names as ethnically German, and students with lower class names as having a middle class background, our estimates of both ethnic discrimination and social class discrimination would be downwardly biased. Also, the estimate of ethnic discrimination would be downwardly biased if teachers perceived students with Turkish names as having a higher social class background than students with ethnic majority lower class names.

To better understand how severe these potential problems really are and how they compare to other studies, consider this: First, my estimates are biased if and only if the numbers above deviate from the perception teachers in the population have. So, if in the population, for instance, the same sizable minority of teachers perceives students with a Turkish immigrant background not as having a Turkish background but as ethnic majority German, then the numbers above do not indicate bias in my

38 Although Justin and Jacqueline are not traditional German names, foreign names are popular among German families with a lower socioeconomic background. Evidently, the teachers in the sample recognize this fact.

estimates. However, I do not know this number. Secondly, the numbers reported abive are similar to those reported for the perception of names of Blacks and Whites in the US: Gaddis (2017a) finds congruent perceptions of 87.3% for first names held by Whites and 75.0% for first names held by Blacks. That doesn't mean that my numbers are fine, but that in other countries and cultures, similar rates and differences are found.

Thirdly, and maybe most importantly, a closer look at the data reveals that the deviations on all three dimensions—ethnicity signal of Turkish names, class signal of lower class names, and class signal of Turkish names—are highly correlated. It is virtually the same group of teachers who deviates on all three items, except for the additional some 20% that declare to perceive the Turkish names as middle or upper class. The behavior of this group of teachers could well be a manifestation of social desirability bias instead of a real difference in perception. Admittedly, I do not know whether the deviations are due to socially desirable behavior by some teachers. I also don't know whether this social desirability bias, should it indeed be the explanation of the pattern I find, may also influence teachers' behavior in the field and, thus, *not* bias our estimates of discrimination.

6.5.4 Questionnaire[39]

Teachers had to answer each item of the online questionnaire and could not go back once they had left a page. This was done to prevent teachers from skipping back and changing their evaluation of the student's performance as a reaction to later items. At the beginning of the survey, teachers were presented with the essay, information about the specifics of the assignment and the information that it was written by a ten year old fourth grader with a particular name. They were then asked to evaluate the child's performance with the following items:

1. "Which grade would you give [name of child] for this essay?" Teachers could assign German grades from 1 (best) to 6 (worst), including plus and minus signs to differentiate further between full grades.

2. "How likely is it that [name of child] can keep up in German lessons at the *Gymnasium* with this performance?", rated on the scale of 1 to 5.

The essay was visible for each of these items. Then, teachers answered a few questions about their teaching experience, the ethnic and social composition of their class,

39 Thanks to Anne Landhäußer for programming the questionnaire.

and their own social and ethnic background. On average, teachers took 12 minutes to complete the survey. For screenshots of all pages of the questionnaire please see appendix D or the OSF project at https://osf.io/dqtkg/.

6.6 Analytic Strategy

6.6.1 Essay Grading

In order to assess discrimination in grading I model the grade as given by the teacher with two different models of which the following is the more simple one:

$$Y_i = \alpha + Q_i\beta_1 + F_i\beta_2 + T_i\beta_3 + L_i\beta_4 + \boldsymbol{C}\boldsymbol{\gamma} + \epsilon_i \tag{6.1}$$

where Y is the grade assigned to essay i, Q captures the quality of the essay ($good = 1$), F identifies the gender of the name attached to the essay ($female = 1$), T distinguishes between Turkish and German names ($Turkish = 1$), and L stands for the social class associated with a German name ($low\ social\ class = 1$). Q, F, T, and L are dummy variables, $\boldsymbol{C}$ is a vector of controls. My coding dictates that the bad essay, male names, and those representing an upper middle class background are the according reference groups. ϵ is an error term with the usual properties in an OLS scenario. To assess the sensitivity of the standard errors, I also estimated models featuring heteroskedasticity-robust standard errors. This did not change the significance levels of any of the parameters.

In the second model I examine interaction effects between some of the variables. This model examines whether name effects depend on the quality of the essay and looks as follows

$$Y_i^* = \alpha + Q_i\beta_1 + F_i\beta_2 + T_i\beta_3 + L_i\beta_4 + (Q_iT_i)\beta_5 + (Q_iL_i)\beta_6 + (Q_iF_i)\beta_7 + \boldsymbol{C}\boldsymbol{\gamma} + \epsilon_i \tag{6.2}$$

Now, β_1 captures the difference between bad and good essay for German upper middle class boys, β_2 captures gender differences in the bad essay, β_3 estimates the difference between Turkish and German upper middle class names when the essay is bad, and β_4 the one between German lower class and upper middle class names for the bad essay. Finally, β_5, β_6, β_7 estimate whether the effects estimated by β_2, β_3, and β_4 are any different when the essay quality is good instead. Obviously, some of the effects of interest the results rely partly on the sums or differences of these coefficients and the corresponding confidence intervals. For example, $\beta_3 + \beta_5$ yields the difference

Table 6.1: Teachers' expectations, dependent on essay quality.

Likelihood of keeping up at the *Gymnasium*	Essay Quality						
	good		bad		Total		
	No.	Col %	No.	Col %	No.	Col %	Cum %
1 (very unlikely)	22	19.5	58	46.8	80	33.8	33.8
2	32	28.3	40	32.3	72	30.4	64.1
3	42	37.2	21	16.9	63	26.6	90.7
4	17	15.0	3	2.4	20	8.4	99.2
5 (very likely)	0	0.0	2	1.6	2	0.8	100.0
Total	113	100.0	124	100.0	237	100.0	

between good essays with Turkish and German upper middle class names on them and $\beta_1 + \beta_5$ estimates the returns to a good essay compared to a bad essay for students with a Turkish name. Ethnic discrimination for the bad essay is returned by $\beta_3 - \beta_4$, for the good essay by $(\beta_3 + \beta_5) - (\beta_4 + \beta_6)$.

The participating teachers graded the essays according to a usual 15 point German grading scale, $Y = \{1, 1-, \ldots, 5-, 6\}$, turned into a scale ranging from 0 (worst grade, German 6) to 14 (best grade, German 1), $Y = \{0, 1, \ldots, 13, 14\}$. Empirically, teachers assigned grades from 2 (German 5) to 12 (German 2+).

6.6.2 Teachers' Expectations

In the German education system, elementary school teachers' expectations about the future development of students' abilities and skills are crucial for students' success in the education system. In fourth grade teachers recommend a secondary school track to each child. Practically speaking, they need to answer the question whether the child in question will be able to keep up at the school tracks offered in a particular state. Since the *Gymnasium* is the highest track available in all federal states, I focus on teachers' estimation of the likelihood that the child can keep up in German lessons at the *Gymnasium*.

Discrimination in expectations is assessed by modeling the probability of having a teacher assigning a particular likelihood of success. This ordinal variable originally has five categories with the endpoints labeled as "very unlikely" (1) and "very likely"

(5), respectively. I model this ordinal dependent variable using an *ordinal logit model* (OLM) (Long, 1997).[40]

For both essays, teachers are hesitant to assign high likelihoods of success (see table 6.1)—in fact, the highest category (5: very likely) was used only twice. In part, this is probably due to the average to low grades teachers have given for the essays—grade and expectations are correlated ($r = 0.51$, $p < .001$). On the other hand, teachers have admittedly little information about the child's true ability after only one short essay, and we know from past research that German elementary school teachers tend to be risk averse when making track recommendations (Maaz et al., 2008). Thus, we should expect them to be cautious in their estimation of children's potential. Due to the skewed distribution, I recoded the variable so that the three highest categories (medium to high likelihood of success) were collapsed into a single category. The dependent variable now has three categories ($J = 3$), and is linked to the measurement model of the OLM as follows

$$Y_i = \begin{cases} 1 \Rightarrow 1 \text{ (“very unlikely”)} & \text{if } \tau_0 = -\infty \leq Y_i^* < \tau_1 \\ 2 \Rightarrow 2 & \text{if } \tau_1 \leq Y_i^* < \tau_2 \\ 3 \Rightarrow 3, 4, \text{ and } 5 \text{ (“very likely”)} & \text{if } \tau_2 \leq Y_i^* < \tau_3 = \infty \end{cases} \tag{6.3}$$

where τ_1 through τ_{J-1} are cutpoints estimated in the OLM (Long, 1997).

As in equation 6.1, I model the underlying latent variable of the OLM as follows

$$Y_i^* = \alpha + Q_i\beta_1 + F_i\beta_2 + T_i\beta_3 + L_i\beta_4 + \boldsymbol{C}\boldsymbol{\gamma} + \epsilon_i \tag{6.4}$$

where Q captures the quality of the essay ($good = 1$), F identifies gender ($female = 1$), T distinguishes between Turkish and German names ($Turkish = 1$), and L stands for the social class associated with a German name ($low\ social\ class = 1$). $\boldsymbol{C}$ is a vector of controls and ϵ is a random error that follows a logistic distribution with mean 0 and variance $\pi^2/3$.

For teachers' expectations I also estimate a model featuring interaction effects using the same model specification as for grading (see equation 6.2):

$$Y_i^* = \alpha + Q_i\beta_1 + F_i\beta_2 + T_i\beta_3 + L_i\beta_4 + (Q_iT_i)\beta_5 + (Q_iL_i)\beta_6 + (Q_iF_i)\beta_7 + \boldsymbol{C}\boldsymbol{\gamma} + \epsilon_i \tag{6.5}$$

40 I also ran all models as OLS regressions using the original Likert scale, as well as logistic regressions using a dichotomized variable combining categories 1 and 2 versus 3 to 5. Substantively, this does not alter my conclusions as discussed below. For results and syntax see the supplementary material at https://osf.io/dqtkg/.

The interpretation of the coefficients is also just like in equation 6.2, only that now the dependent variable represents the underlying latent variable of the OLM. For both model specifications, I test the *parallel regression assumption*, also known as *proportional odds assumption*, using the Wald test suggested by Brant (1990). Results suggest that the assumption holds for both models.

To foster interpretation of the results from this non-linear model and to address problems of group comparisons (Allison, 1999; Karlson et al., 2012; Long, 2009; Mood, 2010), I calculate and plot the probabilities of falling into the different categories of the dependent variable as

$$Pr(y = m|\mathbf{x}) = F(\tau_m - \mathbf{x}\beta) - F(\tau_{m-1} - \mathbf{x}\beta) \tag{6.6}$$

where F is the cdf for ϵ and is logistic with $Var(\epsilon) = \pi^2/3$. In order to calculate discrete change effects in the probability for a specific change in one of the independent variables, I take the difference of two probabilities:

$$\frac{\Delta Pr(y = m|\mathbf{x})}{\Delta x_k} = Pr(y = m|\mathbf{x}, x_k = 1) - Pr(y = m|\mathbf{x}, x_k = 0) \tag{6.7}$$

6.6.3 Analysis Sample

All models are estimated using a restricted sample: I only look at teachers who report to have students with an immigrant background in their classes. The reason for this is, once again, my concern about external validity. Following arguments from labor economics (Becker, 1957/1971; Heckman, 1998) discussed above in this chapter as well as elsewhere in this study, I posit that teachers' behavior toward ethnic minorities only matters as long as they actually teach them and, hence, have the opportunity to discriminate against them. Thus, by restricting the sample to those teachers who do teach children of immigrant background, I arrive at a more accurate estimation of discrimination in the actual school context. I lose a few further cases because I control for background variables, of which some have missing data. The variables I control are the teacher's sex, immigrant background, and teaching experience, as well as the education of the teacher's parents. This way, the sample shrinks to $N = 199$.[41]

41 I also ran all models with the full sample, with similar results and substantively unaltered conclusions. Results and Stata syntax for replication purposes are available at https://osf.io/dqtkg/.

Table 6.2: Summary statistics of grades, dependent on child's name and essay quality.

	N	Mean	SD	Median	Min	Max
Good essay						
Jakob	18	6.72	1.87	7	3	10
Sophie	16	7.56	2.16	8	4	10
Justin	21	7.43	1.94	7	4	11
Jaqueline	21	6.90	1.61	6	5	12
Murat	19	6.95	2.46	8	2	11
Ayse	18	7.83	1.76	8	5	12
Total	113	7.22	1.97	7	2	12
Bad essay						
Jakob	24	5.71	2.14	6	2	11
Sophie	22	5.45	2.06	6	2	10
Justin	12	4.83	1.40	5	2	7
Jaqueline	18	6.11	1.68	6	3	9
Murat	19	5.16	1.89	5	2	9
Ayse	29	5.69	1.81	5	3	9
Total	124	5.55	1.88	6	2	11

6.7 Results[42]

6.7.1 Grading

The essay that was pretested as "good" received rather average grades, with a mean of 7.22 ($SD = 1.97$). However, it is significantly better than the bad essay that has a mean of 5.55 ($SD = 1.88$; $t = 6.68$, $p < 0.001$). Table 6.2 shows basic summary statistics for each of the six names, again separated for the good and bad essay. Although there is some variation, no clear patterns are visible. In fact, there are no significant differences in the mean grade between child names. Apparently, the name of the child does not have a systematic impact on essay grading. This conclusion is also supported by

42 Results and Stata syntax for replication purposes is available at https://osf.io/dqtkg/.

Table 6.3: Regression of essay grades on essay quality, child's gender and child's background.[a b]

		Model 1.1		Model 1.2	
Essay quality: good	β_1	1.65^{**}	(0.27)	1.50^{**}	(0.55)
Gender: female	β_2	0.38	(0.27)	0.43	(0.38)
Name: Turkish[c]	β_3	-0.10	(0.33)	-0.28	(0.44)
Name: German lower class[c]	β_4	0.14	(0.34)	0.08	(0.49)
Turkish × good quality	β_5			0.41	(0.67)
Lower × good quality	β_6			0.16	(0.68)
Female × good quality	β_7			-0.08	(0.54)
Constant		5.95^{**}	(0.44)	6.02^{**}	(0.47)
Observations		199		199	
R^2		0.211		0.213	

† $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

[a] Unstandardized coefficients; standard errors in parentheses.
[b] The model includes controls for teacher characteristics (gender, parental education, migration background, years of teaching experience)
[c] Reference group: German upper middle class name.

the regressions I ran, as can be seen in table 6.3. Except the coefficient for essay quality, β_1, no coefficient turns out significant on conventional levels. The same holds for the linear combinations that allow to test whether grades differ between groups for the good essay. The results also clearly show that there are no group differences in returns to the good essay compared with the bad essay—all interaction effects with essay quality are far from being statistically significant on conventional levels. Thus, I find no evidence of discrimination in essay grading and reject the hypotheses 1_a, 1_b, and 1_c. However, the results are perfectly compatible with hypotheses 2_a, 2_b, and 2_c.

6.7.2 Expectations

In contrast to the results for grades, I do find a significant effect of a student's background in the expected directions in the ordinal logit model (see table 6.4, model 2.1): with the performance shown, children whose name indicates a Turkish background are perceived to be less likely ($\beta_3 = -.94$, $p < .01$) to succeed at the *Gymnasium* than children with a German upper middle class background (reference category). The dif-

Table 6.4: Ordinal logistic regression of expectations on essay quality, child's gender and child's background.[ab]

		Model 2.1		Model 2.2	
Essay quality: good	β_1	1.17^{**}	(0.28)	2.03^{**}	(0.56)
Gender: female	β_2	0.03	(0.28)	-0.02	(0.28)
Name: Turkish[c]	β_3	-0.94^{**}	(0.34)	-0.35	(0.43)
Name: German lower class[c]	β_4	-0.52	(0.35)	-0.31	(0.48)
Turkish × good quality	β_5			-1.55^{*}	(0.72)
Lower × good quality	β_6			-0.76	(0.74)
Female × good quality	β_7			-0.40	(0.56)
τ_1		$-0.87^{\dagger}$	(0.45)	-0.64	(0.48)
τ_2		0.54	(0.45)	0.79	(0.48)
Observations		199		199	
Log Likelihood		-203.46		-201.02	
Pseudo R^2 (McFadden)		0.07		0.08	

† $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

[a] Unstandardized coefficients; standard errors in parentheses.
[b] The model includes controls for teacher characteristics (gender, parental education, migration background, years of teaching experience)
[c] Reference group: German upper middle class name.

ferences between German names signaling different social classes ($\beta_4 = -.52$, $p > .1$) and between Turkish names and German lower class names ($\beta_3 - \beta_4 = -.42$, $p > .1$) are in the expected direction, but not statistically significant. This simple model contests hypotheses 2_a and 2_b. However, note that if I drop L from equation 6.5 and, thus, compare the results for Turkish names to all German names no matter what class connotation they have, I get $\beta_{3^*} = -.68$ ($p < .05$). This result suggests that students with a Turkish background suffer from discrimination at least partly because they are from a lower social class. Given the insignificant difference to German lower class names, ethnic discrimination alone does not seem to be the decisive factor. Also, this first model shows no evidence for discrimination based on gender—the corresponding coefficient is virtually zero. This result clearly rejects hypothesis 2_c.

Next, I added interaction effects between essay quality and child's background (table 6.4, model 2.2) to investigate different returns to performance for the three groups.

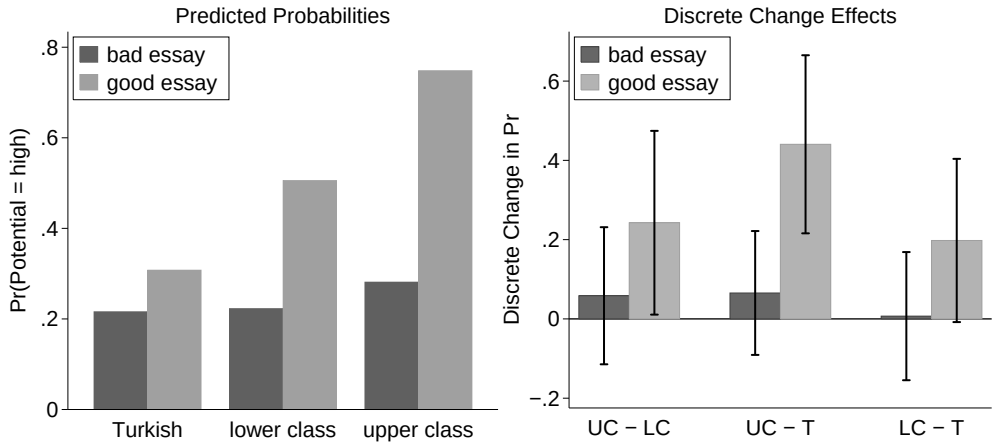


Figure 6.2: Left panel: Predicted probabilities for a high likelihood of success at the *Gymnasium*, dependent on name and essay quality. All other variables held constant at the mean. Right panel: Discrete changes in probabilities for each of the three contrasts—upper class vs lower class (UC – LC), upper class vs Turkish (UC – T), lower class vs Turkish (LC – T), with 95% confidence bars. Calculations are based on model 2.2 in table 6.4.

It turns out that the previously discovered advantage for the German upper middle class depends on essay quality. If the essay is bad, there is no significant difference in estimated expectations for either contrast between the three groups ($\beta_3 = -.35$, $p > .1$; $\beta_4 = -.31$, $p > .1$; $\beta_3 - \beta_4 = -.04$, $p > .1$). However, for the better essay, German upper middle class children have—on the 10% level—significantly higher odds than German lower class ($\beta_4 + \beta_6 = -1.07$, $p = .059$) and Turkish children ($\beta_3 + \beta_5 = -1.9$, $p = .001$) to be trusted to succeed at the *Gymnasium*, and German lower class children in turn have higher odds—significant on the 10% level—than Turkish children ($(\beta_3 + \beta_5) - (\beta_4 + \beta_6) = -.83$, $p = .067$). Thus, I find no evidence for discrimination for the bad essay, but I do find evidence for discrimination on the basis of social class, as evidenced by the contrast between German upper and lower class names, and ethnicity, captured by the contrast between Turkish and German lower class names. These results suggest that hypotheses 2_a and 2_b hold only conditional on essay quality.

Another interesting result is that in this model, essay quality is not a significant predictor of the teachers' expectation towards Turkish children ($\beta_1 + \beta_5 = .48$, $p > .1$), whereas it does matter for both groups of German children (lower class: $\beta_1 + \beta_6 = 1.27$, $p < .01$; upper middle class: $\beta_1 = 2.03$, $p < .01$). Put differently, the returns to performance are not statistically different from zero and therefore lowest for the Turkish students, higher—albeit not significantly ($\beta_6 - \beta_5 = .79$, $p > .1$)—for German

lower class children, and highest for upper middle class children, whose returns are significantly higher than those of Turkish children ($\beta_5 = -1.55, p. < .05$).

Predicted probabilities as effect sizes

To get a more vivid impression of the effect sizes behind the coefficients from the otherwise hard to interpret ordinal logit model, I visualize the results of model 2.2 in figure 6.2. On the left panel it shows predicted probabilities for falling into the category with the highest likelihood of success at the *Gymnasium* as assigned by the teachers for the three groups of students whose contrasts define social class discrimination and ethnic discrimination. The right panel shows discrete change effects in the corresponding probabilities for the three contrasts.

From figure 6.2, the difference between the two essays becomes very clear: for the bad essay, it is simply not important who wrote it—the slight advantage for German upper middle class children is clearly not significant. Among those who supposedly wrote a good essay, group differences increase substantially: Teachers assign significantly higher probabilities to children with a German upper middle class name compared to children with a German lower class name and children with a Turkish name. The difference between children with German lower class names and those with a Turkish name is of about an equal magnitude as that between the two German groups and narrowly misses the 5% significance level.

6.8 Discussion

In this chapter, I have presented the design and the results of an experimental study that was explicitly designed to address common shortcomings of previous experimental research on discrimination in education. Shortcomings I sought to address concerned the usage of biased or uninformative samples, ignored effect heterogeneity across the distribution of observed performance, and, last but not least, the confounding of ethnic and social discrimination. To address these issues I set up a $2 \times 2 \times 3$ factorial design, varying essay quality, gender, and social and ethnic class background. I examined discrimination by teachers in two outcomes in German elementary school that are of critical importance for children's educational achievement: grades and track recommendations. I argued that examining discrimination in these different outcomes and situations also allows to indirectly test different theories of discrimination.

Findings

Different regression models of *grades* on essay quality and students' names show no evidence of discrimination by virtue of students' ethnic or social class background or gender. In contrast, I find statistically significant differences in teachers' *expectations* between German upper middle class and Turkish students averaging over the competence distribution. The estimate for ethnic discrimination, the difference between Turkish and German lower class students, is not statistically significant. However, if I do what other studies typically do—i.e., if I lump together German students of different class background—I find a statistically significant difference between German students and Turkish students. Thus, it seems that social class discrimination plays a major and probably more important role in discrimination against Turkish students than previously thought on the basis of experimental studies that do not disentangle ethnic from social class discrimination.

A second model featuring interaction effects reveals that there is indeed discrimination on the basis of both social class and ethnicity but only for the better of two essays that, in fact, turned out to be rather average, even though it was pretested as "good" in a federal state with comparable standards. The results from the model with interaction effects can also be interpreted as evidence for differential returns to essay quality: for Turkish students the returns are lowest, followed by German lower class students, followed by German upper middle class students for whom returns are highest.

A very clear and robust finding over all models for both outcomes is that teachers do not discriminate on the basis of gender.

Implications for theories of discrimination

Taken together, the results provide evidence against more simple models of ingroup-favoritism or outgroup derogation, such as social identity theory (Tajfel, 1982; Tajfel & Turner, 1986). For those who think that Becker (1957/1971)'s model of taste discrimination is applicable to either or both situations I investigated, the results would also provide evidence against this model. However, I have argued that the mechanism Becker (1957/1971) suggests is not really applicable to either situation, so that I think that the results should not be read as evidence against Becker (1957/1971). Also, his theory might be helpful in understanding discrimination in other situations in education.

The other models I have discussed and the hypotheses derived from the models receive more support from the findings. A statistical discrimination model with group-

specific reliabilities (Aigner & Cain, 1977)—lowest reliability for Turkish students, followed by German lower middle class, and upper middle class students—is in line with the findings of different returns and, thus, the observed interaction with essay quality. Such a model may also feature risk-averse teachers (Maaz et al., 2008) and stereotypes that are biased to the disadvantage of Turkish students and students from lower social class families without immigrant background (see chapter 5).

The findings also appear to be in line with the mechanism proposed by the continuum-model by Fiske et al. (1999), Fiske and Neuberg (1990). The good essay that turned out to be rather average might have been not bad enough to move teachers from a category-based response based on stereotypes to a piecemeal-integration of individuating information in case of upper middle class students. Conversely, it might have not been good enough to foster the same process for the Turkish students and they, too, were treated according to the teachers' stereotypes. However, since the bad essay turned out to be really bad, teachers might have turned from a category-based judgment to a more individuating judgment in case of upper middle class students and, thus, have graded them as bad as Turkish and lower class students. Similarly, based on aversive racism theory we could also explain the observed pattern: In case of the supposedly good but really rather average essay, teachers might have taken advantage of the ambiguity of the situation and treated students according to their stereotypes and prejudices.

One possibility to distinguish between different theoretical mechanisms might be to assess discrimination additionally for an excellent essay. While models of statistical discrimination that rely on group-specific reliabilities would predict an even larger gap between students' from different social and ethnic groups, applying the mechanisms from the continuum model or aversive racism theory probably leads to the opposite prediction of less discrimination on the basis of these factors.

Individual versus group discrimination

I interpret the group differences found in the present study as individual discrimination. Depending on the exact contrast I find individual discrimination by virtue of social class and ethnicity. I also suggest that if teachers recommend tracks according to the pattern observed here, group discrimination should occur. As discussed at several points in this dissertation (e.g., in section 3.1.2), models of statistical discrimination that predict the observed pattern also explain group discrimination—depending on the exact model for both categorical and continuous outcomes or categorical outcomes only (Aigner & Cain, 1977). From the less formalized continuum model and

aversive racism theory it is less clear to make such a prediction, but should teachers follow the mechanisms suggested by these theories it would be difficult to explain how—on the group level—the effects found on the micro level should disappear.

Whether the findings are also indicative of group discrimination in the real world, depends, of course, also on questions of external validity and how good of a proxy the expressed expectations are for actual track recommendations. I discuss these points in some more detail below.

Will the real ethnic discrimination effect please stand up?

I have argued that experimental studies typically confound discrimination on the basis of ethnicity or immigrant background or race and social class or socioeconomic background. However, note that this position is not immune to critique. In section 2.3.3 I have suggested that ethnic discrimination is the *total* causal effect of an ethnic information about or an ethnic signal sent out by an individual on how this individual is treated by another individual.

I see two different but related lines of reasoning that could be brought forward against the strategy implemented in my experimental study. First, should social class really be held constant when examining ethnic discrimination? When, as I have argued, social class and ethnic signals are interpreted as confounders, identifying ethnic discrimination indeed requires to hold constant social class. However, one might argue that the social class content of an ethnic signal is merely mediating a part of the total effect of the ethnic signal and, thus, should not be held constant if interest lies in the total causal effect of the ethnic signal. Secondly, one might argue the other way around, namely that if controlling for a social class signal is said to be necessary to identify ethnic discrimination, why then is it not necessary to control for all other confounding signals?

I have no final answer to these questions that I could offer here. However, the answer will certainly depend on our understanding of what a signal or an information is and how these should be distinguished from the beliefs and attitudes they trigger in the mind of the perceiver. Certainly, signals and information occur prior to cognitive processes in the perceiver's mind that handles them. Since many supposedly confounding signals might not be signals after all but rather contents of—possibly biased—stereotypes and attitudes, we would not want to control for them if our interest lies in the total causal effect of a signal or an information.

From this perspective, a solution and answer to the question raised above might be to hold constant all information and signals perceivers have at their disposal in the

real-world situation under study but to *not* hold constant information that perceivers do not have access to and, thus, can only fill in by a process of stereotyping or applying prejudices. Correspondence studies on labor market discrimination, for example, follow this approach and send out applications that are no more or less informative than other applications or the paired application. Thus, they hold constant all the information an employer has access to—but not more. Of course, information typically not available to the perceiver could nevertheless be of diagnostic value for the outcome under study. However, to identify individual discrimination, controlling for information the perceiver does not know but only has stereotypic knowledge about would mean to induce an overcontrol bias to the estimate of the total causal effect that also conceptualizes individual discrimination arising from statistical discrimination as discrimination. But, to identify group discrimination more directly than in the present study and typical correspondence studies, controlling for the information teachers do stereotype about seems necessary.

While these questions and considerations might appear to be nit-picking, the answers to them could be of great relevance for all doing research on discrimination. Note, however, that the questions raised here are partly methodological, partly theoretical questions. Only the methodological questions may be answered by a definition. One thing I feel rather safe to conclude from this discussion and the discussion in chapter 2 is that researchers should be as clear as possible with regard to the meaning of terms such as discrimination or ethnic discrimination in particular and also with regard to their identification and estimation strategy.

6.9 Limitations and Directions for Future Research

The experiment presented in this chapter was designed to address various shortcomings and desiderata of previous experimental research in education—not only but especially in Germany. Of course, it has itself several limitations that I will briefly discuss in this section. Addressing these limitations will be a task for future research.

External validity

To assure a high external validity, I have sampled teachers instead of students and the response rate turned out to be much larger than in comparable studies (e.g., Sprietsma, 2013). I also analyzed only responses of teachers that actually teach immigrants in their classes. However, the external validity of a study also hinges upon how real-

istic and lifelike the experimental situation is. In this regard, my experiment is closer to a typical lab experiment than to a field experiment.

Statistical power

$N = 237$ teachers in the whole sample and $N = 199$ in the analysis sample provided not enough statistical power for investigating fully interacted models for the $2 \times 2 \times 3$ factorial design. It would have been interesting to also examine interaction effects of ethnic and social class background with gender. This should be addressed by future research on the backdrop of findings about differential attitudes towards boys and girls with immigrant background and gender inequalities in education among immigrants more generally (Fleischmann et al., 2014; Glock & Klapproth, 2017). More statistical power is also needed to investigate the effects of classroom and teacher characteristics.

Discrimination? In track recommendations?

I have theorized about discrimination in track recommendations more generally and argued that expectations determine not only track recommendations but also many other important decisions in education such as ability grouping within tracks and may turn into self-fulfilling prophecies. However, it is unfortunate that I have not explicitly asked teachers which track they would recommend based on the observed performance. I can only hypothesize about potential differences to my findings for expectations of future performance. Discrimination in an outcome explicitly asking for track recommendations would have probably been higher, since variables such as parental support and involvement should play an even more important role than for the more narrow question on future performance in one subject. If anything, the more narrow question should reduce the effect of students' social and ethnic background and render my estimate of discrimination conservative. However, this remains speculative unless empirically addressed in future research.

Also, the question I asked and examined is the result of either stereotyping or, in case of aversive racism, maybe applied prejudice. Neither is it actual behavior—except for ticking a box in the questionnaire—nor is it a behavioral intention. Here, too, the question which track the teacher would recommend, would have been a very interesting outcome to look at.

Mechanisms of discrimination

As discussed above, the present experiment only provides indirect evidence about theoretical mechanisms of discrimination in education. For a more direct test of different theories of discrimination and their proposed mechanisms, direct measures of, for example, stereotypes and prejudices would be needed. Unfortunately, the Ministry of Education in Baden-Württemberg did not approve of such items and I had to drop them. Future research should seek to implement such measures. However, future studies investigating the mechanisms of discrimination in education may also build on and adapt the research design of the present study without introducing measures of stereotypes or prejudices. An example for an indirect test using an excellent essay, I have given above.

Classroom and teacher characteristics

In a short questionnaire after collecting the data on grades and expectations, I also asked teachers to answer a few questions on classroom characteristics such as the proportion of immigrants in the class and the social background of students as well to report some personal demographics, namely their age and gender, their parents' education, their immigrant background, and years of teaching experience. An explicit look at the effects of these variables was beyond the scope of the present study. From some preliminary results I can tell that most of these variables have no statistically significant effects on discrimination in grades or expectations. I found an effect for teachers' work experience—or, alternatively, for the highly correlated variable age—that suggests that less experienced (i.e., younger) teachers discriminate mainly on the basis of social class and less so on the basis of immigrant background. More experienced teachers show the exact opposite pattern. Whether this effect holds in studies with more statistical power and how to theorize the effect of different classroom and teacher characteristics, I have to leave to future research.

7 Conclusion

7.1 What Have We Learned?

My aim in the present study was to broaden our knowledge regarding discrimination in education by making methodological, theoretical, and empirical contributions. I was concerned primarily with *ethnic* discrimination, followed by *social class* discrimination, and *sex* or *gender* discrimination.

Two motives for studying discrimination

I have argued that there are two major reasons why we study discrimination and that these reasons are related to different definitions or forms of discrimination. First, discrimination by virtue of characteristics such as race, ethnicity, social class, or gender is typically considered unfair or unjust by most people in developed countries. Thus, discrimination may be studied in its own right, that is, it may simply be the explanandum in an analysis. For such a perspective it might suffice to look at individual discrimination, that is, discrimination as individual-level causal effect.

However, secondly, studies of discrimination are often motivated by inequalities between different ethnic and social groups. Studying discrimination as an explanation for disparities between groups makes it necessary to move beyond the explanation of discrimination as individual-level causal effect and to also examine group discrimination. Because individual discrimination does not necessarily aggregate to group discrimination, it is necessary to address the difference between these forms of discrimination properly when defining, identifying, and estimating discrimination.

Definitions of discrimination

As a methodological contribution, I have discussed several different definitions of discrimination to find a logically consistent and useful one. I have argued that discrimination in general is best understood as the causal effect of an information about or a signal sent out by an individual on how this individual is treated by another individual. In the the present study I was mainly interested in ethnic discrimination that I—based on the general definition—defined as the causal effect of an ethnic information about or an ethnic signal sent out by an individual on how this individual is treated by another individual.

I have argued that my general definition is the most useful starting point for defin-

ing more concrete forms of discrimination, for different reasons: The treatment, an information or signal, is truly manipulable and allows to ask well defined causal questions based on meaningful alternative causal states. That is, the definition avoids the problem of defining discrimination as the total causal effect of immutable characteristics assigned early in life that leads to an undesirable conflation of unconditional inequality with discrimination. Furthermore, it circumvents issues that arise when discrimination is defined as a direct effect. Last but not least, it avoids vague terms that are hard to define or constraints of the phenomenon of discrimination that are hard to justify—both usually carry normative connotation.

Recall that I have not concluded from this that discrimination should be only understood as behavior that leads to inequality on the group level. This only leads to a very narrow understanding of discrimination and leaves us with the problem of finding different terms for particular types of discrimination that—under certain conditions—do not lead to inequality between groups. Statistical discrimination, for example, is discrimination by all means of a useful understanding of the term. That it does not lead to inequality on the group level under all—but certainly some—circumstances does not make it less discriminatory. To acknowledge that individual discrimination and group discrimination are not the same thing remains of great importance, nevertheless, or, maybe, because of the more general nature of the definition proposed in this study.

Last but not least, based on the design of my experimental study, I came to realize that my definition does not solve all methodological problems without further discussion. The questions what are treatments and what are potential confounders and what are mere mediators in a model of discrimination based on my general definition are crucial questions. Answers to these questions require both methodological and theoretical input. I will briefly return to this point below when I recap the results of the experiment.

Theories of discrimination

As a theoretical contribution, I have discussed several different theories of discrimination from different disciplines and applied them to typical situations at the end of German elementary school. I have argued that, contrary to a popular line of reasoning, Becker (1957/1971)'s theory of taste discrimination is less applicable to education in general and to key situations in German education in particular. Conversely, I have argued that models of statistical discrimination (e.g., Aigner & Cain, 1977) or models that built on the statistical discrimination mechanism, such as error discrimination

(e.g., England & Lewin, 1989) or inaccurate statistical discrimination (Bohren et al., 2019), are indeed more useful than sometimes suggested. I have argued that they are applicable to many situations including the situation of recommending tracks at the end elementary school and that there are several models of or related to statistical discrimination that are able to explain group discrimination and, thus, inequality.

I have criticized institutional discrimination approaches as not very helpful to understand discrimination, since they lack clear causal mechanisms on both macro- and micro-level. Much more useful are three different theories from social psychology that have I discussed: Social identity theory (SIT; Tajfel, 1982; Tajfel & Turner, 1986), the continuum model (Fiske et al., 1999; Fiske & Neuberg, 1990), and aversive racism (Dovidio & Gaertner, 2008; Gaertner & Dovidio, 1986). They all provide micro mechanisms that are applicable to education and may help to understand discrimination by teachers.

From the theories I deemed useful and applicable, I later derived hypotheses that I tested in my experimental study. My discussion was also meant to show how important it is to understand the key determinants of discrimination—prejudices and stereotypes—both of which I investigated in the following chapters.

Teachers' prejudices and stereotypes

Research on teachers' *prejudice* often relies on geographically limited convenience samples of students. Using data from the German general social survey, ALLBUS, I have quantified the bias in one of those studies and argued that, on the backdrop of the size of the bias, more representative research is needed. To this end, I have used ALLBUS data and have shown that teachers in German education hold negative prejudices about Turks but less negative prejudices about Eastern Europeans of German descent and virtually no negative prejudices about Italians.

I have introduced a new instrument to investigate teachers' *stereotypes* about the average competences in math and reading of different groups of students. Comparing teachers beliefs to actual group differences in published studies shows that teachers correctly rank different groups of students, that is, teachers know which groups perform better or worse than other groups. However, I also find that teachers' stereotypes are probably biased to the disadvantage of boys, students from lower social class families or—in comparison with students from upper class families—students from middle class families, immigrants in general, as well as immigrants of Turkish and Russian origin in particular. These results are in line with theoretical predictions and, thus, I have argued that they speak to the validity of the new instrument.

The findings that teachers hold both negative prejudices against some but not all ethnic groups and biased stereotypes about different groups of students are of great importance for a better understanding of discrimination in education. The findings suggest that, once individual discrimination is established, it is rather likely to also aggregate to group discrimination and, thus, help to explain inequality.

Disentangling ethnic from social class discrimination

To address the question of discrimination in education empirically, I set up an experimental study that investigated discrimination in grading and teachers' expectations about future performance by virtue of ethnic and social class background as well as gender.

To address shortcomings of prior studies, I drew a random sample of elementary schools from a German federal state and, in my analysis, focused on the responses of teachers that actually teach students of immigrant background in their classes to enhance external validity. I varied essay quality using two different essays of which one was pretested as bad and one as good to assess discrimination at different points in the distribution of observed performance to investigate group differences in returns or reliabilities. The main methodological and substantive contribution of my study was the attempt to disentangle ethnic discrimination from social discrimination; prior experimental studies that made use of names or photos as stimuli ignored that ethnicity and social class and, thus, ethnic discrimination and social class discrimination, are very likely to be confounded in virtually every society. I have argued that by ignoring social class as a confounder, estimates of ethnic discrimination are upwardly biased, that is, they overestimate the part ethnicity plays in discrimination a particular ethnic group may suffer from.

The solution I propose is to compare Turkish names to German names with a similar social class connotation instead of a representative selection of German names that would be associated with higher social class background or socioeconomic status, respectively. This way I was able to hold constant social class in my comparison of teachers' responses to allegedly Turkish and German students. This difference, I suggest, may be interpreted as ethnic discrimination. In my analysis I do not find any evidence for ethnic, social class, or gender discrimination in grading—teachers discriminate on the basis of essay quality only. However, my analysis of expectations provide evidence for the suspected confounding of ethnic and social class discrimination. While, in a simple model averaging over both essays, teachers' expectations for Turkish students are not significantly different from the expectations for German

lower class students, they are significantly different from expectations for both upper middle class students and, most importantly, German students *overall*.

A more complex model reveals an interaction effect of group differences with the quality of the essay: For the worse essay no group differences are found. However, for the better essay—that turned out to be rather average—, teachers' expectations differ between all possible contrasts for the three ethnic and social groups. Again, gender plays no role whatsoever in determining teachers' expectations. In sum, there is evidence for discrimination based on both social class and ethnicity. Of course, this result implies that a comparison of expectations for Turkish students and German students overall—as commonly calculated in experimental studies that are not based on extensive vignettes—would yield higher estimates of discrimination.

The following is what I offer as a more general conclusion from my attempt to disentangle ethnic discrimination from social class discrimination in an experiment—even though these insights are not entirely new (see chapter 2 again): Even if experiments are legitimately considered to be the gold standard of causal analysis, by no means do they solve all the problems of causal inference automatically. One problem an experiment cannot solve is the fine articulation of causal states, that is, the precise definition of the causal effect of interest. It is the responsibility of the researcher to make sure that the alternative causal states are described in sufficient detail and that their difference captures what the researcher is substantively interested in—not more, but also: not less. Another and related problem that is not solved by experimental designs, is the problem of theorizing a phenomenon by linking cause and effect through mechanisms.

Mechanisms of discrimination in education

The experiment I conducted also sheds light on the mechanisms of discrimination in education. More simple models of ingroup-favoritism or outgroup-derogation (SIT; Tajfel, 1982; Tajfel & Turner, 1986) cannot explain the observed pattern of no discrimination in grading but discrimination in expectations conditional on essay quality. Models that incorporate situational moderators like imperfect information and ambiguity such as statistical discrimination theory (Aigner & Cain, 1977; Arrow, 1973; Phelps, 1972), the continuum model (Fiske et al., 1999; Fiske & Neuberg, 1990), and aversive racism (Dovidio & Gaertner, 2008; Gaertner & Dovidio, 1986) fare much better in explaining the results. However, distinguishing between these models was not possible using the indirect test based on teachers' responses to the two essays of different quality.

7.2 Where Do We Go From Here?

Even though we have learned quite a bit about discrimination in German education since Kristen (2006b)—the first quantitative empirical study that explicitly theorized discrimination in German education—there is still a lot we would like to learn more about. In the remainder, I would like to stress some points that I think should be addressed and taken into account in future research on discrimination in German education and other countries.

Methodological rigor

After my discussion in chapter 2, I hope it goes without saying that future research should pay more attention to clear and useful language when investigating discrimination. This applies first and foremost to the definition of discrimination. Researchers should be clear about what they mean when they say discrimination and why they study it. This way, the appropriateness of the research design can be examined better than for many past studies. However, even when using a more useful definition than past research, which I suggest I have done in this study, some methodological problems may remain—not to speak of theoretical problems that cannot be solved by a definition but only by theory building and proper application.

It's social class, stupid!

Discrimination by virtue of social class—also known as classism—and the role social class plays in ethnic or racial discrimination should be explicitly addressed by theorizing, identifying, and estimating it using different methodologies and techniques. By this I mean much more than controlling for it in a regression model, which is what is usually done. However, I have shown that experimental evidence on ethnic discrimination in education may to a large part be driven by social class differences between different ethnic signals such as names but, presumably, also photos and other signals or information. Controlling for social class differences in experimental research is usually *not* done. While there might be arguments for why researchers do not want to separate ethnic from social class discrimination, future research should explicitly discuss the problem—irrespective of the particular strategy pursued. In any case: It would be an important methodological contribution to also compare the treatment of Turkish upper middle class children to their German counterparts without falling back on vignettes with explicit and artificial descriptions.

And gender?

One result of my experiment was that gender does not seem to play any role in discrimination in education. However, the sample of my experiment was not large enough to examine interaction effects of gender with both social class and, especially, ethnicity. There is evidence that such interaction effects exist (Glock & Klapproth, 2017), but more research—especially experimental—on these questions is clearly needed.

Theorizing discrimination in education

Especially economists often discuss theoretical mechanisms of discrimination in education rather superficially (e.g., Kiss, 2013; Sprietsma, 2013; van Ewijk, 2011). What is needed is quite the contrary, namely more rigor in theorizing discrimination in education. That does not mean that we should be overly restrictive in applying theories to education. However, if key mechanisms or assumptions of a theory clearly do not apply to a situation or an environment, it might help to remember that there are many models and theories from various disciplines that can help to understand discrimination in education better. For example, given that the findings of the experiment were in line with the continuum model, it might well be worth looking into other dual-process models (see Gawronski & Creighton, 2013, for an overview). The formalized and general model of frame selection (Esser, 2001; Kroneberg, 2010; Kroneberg et al., 2010) might also provide valuable insights. In any case, future research should discuss and rigorously test different theories of discrimination empirically.

Micro level determinants of discrimination

Knowing more about the determinants of discrimination in education should be of great importance. Both implicit and explicit attitudes and beliefs—that is, prejudices and stereotypes—should be studied using unbiased samples of teachers and applying different methodologies. Methods and techniques of data collection including item selection should be guided mainly by theory and the results of prior studies. Other micro level determinants of discrimination are teacher characteristics including personality traits such as social-dominance orientation or right-wing authoritarianism (Altemeyer, 1981; Sidanius & Pratto, 1999; Whitley, 1999).

Macro level and institutional level determinants of discrimination

Theorizing and testing the causal effects of variables on, above, and beyond the classroom level offers the possibility to indirectly test theories and provide evidence with policy implications. Here, I also see potential for developing a useful institutional discrimination approach to education. Of course, such an approach will always need a micro-foundation of which plenty exist in various disciplines.

Discrimination in longitudinal perspective

A perspective I have largely ignored in this study is a longitudinal or dynamic perspective on discrimination. While I have repeatedly pointed to self-fulfilling prophecies and have looked at their major determinants, namely stereotypes, I have not spent much time on discussing other forms of discrimination that take into account time as an important variable, such as cumulative discrimination (Blank et al., 2004). We know that such a dynamic perspective is of great relevance in education (e.g., Jussim, 1989; Jussim et al., 1996; Jussim & Harber, 2005; Jussim, Robustelli, et al., 2009; Lorenz et al., 2016) and, therefore, should be pursued further.

Experiments in lab, field, and conducted by nature

Especially on discrimination in the German education system, more experimental research is needed. For all lab or lab-like designs, such as the one applied in the present study, samples that allow inference to larger populations of actual teachers are vital. While field experiments are certainly more difficult to conduct in education than for example in the labor or housing market, future research should aim at utilizing experimental designs that are more realistic and, thus, provide more direct and unbiased evidence about discrimination by teachers. A first step in this direction could be better cover stories that are realistic and suggest higher stakes. A research design that, if done right, typically features both high internal and high external validity is a natural experiment. Certainly, such a design requires a "natural" treatment, such as a policy change or truly comparable blind versus non-blind scores, for example.

A Items measuring prejudice in Hachfeld et al. (2011)

The items used in Hachfeld et al. (2011) to measure prejudice are taken from the German General Social Survey (ALLBUS). The response scale is a 5-point *agree–disagree* scale ranging from 1 to 5, with higher scores reflecting more prejudiced views toward foreigners (Hachfeld et al., 2011, p. 992). Hachfeld et al. (2011, table 4) report a mean of 1.76 and an SD of $.57$.

Translated items (items 1–3: my own translation; translation of item 4 taken from Hachfeld et al. (2011, table 4)):

1. Foreigners living in Germany should adapt their lifestyle a bit better to that of the Germans.
2. When jobs become scarce, foreigners living in Germany should be sent back to their home countries.
3. Foreigners living in Germany should be prohibited any political action in Germany.
4. Foreigners living in Germany should seek their spouses within their own ethnic group.

The original items are (in German):

1. Die in Deutschland lebenden Ausländer sollten ihren Lebensstil ein bisschen besser an den der Deutschen anpassen.
2. Wenn Arbeitsplätze knapp werden, sollte man die in Deutschland lebenden Ausländer in ihre Heimat zurückschicken.
3. Man sollte den in Deutschland lebenden Ausländern jede politische Betätigung in Deutschland untersagen.
4. Die in Deutschland lebenden Ausländer sollten sich ihre Ehepartner unter ihren eigenen Landsleuten auswählen.

B ISCO-88: Teachers

ISCO-88 unit	Occupations	n	School Teachers	All educators
1210	Directors and Chief Executives (e.g., university chancellor)	3		
1229	Production and Operations Department Managers, n.e.c. (e.g., university president)	0		
1319	General Managers, n.e.c. (e.g., headmaster, school principal)	5		
2300	Teachers (tertiary degree, no further specification)	30	✓	✓
2310	College, university, and higher education teaching professionals	3		✓
2320	Secondary education teaching professionals	14	✓	✓
2331	Primary education teaching professionals	7	✓	✓
2332	Preprimary education teaching professionals	3		✓
2340	Special education teaching professionals	0	✓	✓
2351	Education methods specialists (e.g., curricula developer)	0		
2352	School inspectors	0		
2359	Other teaching professionals, n.e.c.	1		
3300	Teachers (no tertiary degree, no further specification)	1		✓
3310	Primary education teaching associate professionals	0		✓
3320	Pre-primary education teaching associate professionals	17		✓
3330	Special education teaching associate professionals	1		✓
3340	Other teaching associate professionals (e.g., driving instructors)	6		
3460	Social work associate professionals (some work in schools/education)	17		
5131	Child-Care Workers (e.g., nanny)	8		
		$\sum n_j$	51	76

n.e.c.: not elsewhere classified; Sources: International Labour Organization (1990), Geis (2011).

C Measuring Teachers' Stereotypes: Original Instruments

Die NEPS-Studie „Bildungsverläufe in Deutschland“ erfasst die Kompetenzen der Kinder in unterschiedlichen Bereichen. Was denken Sie, wie Schülerinnen und Schüler der zweiten Klassen aus verschiedenen Gruppen im Kompetenzbereich Mathematik abschneiden werden?

Im Vergleich zu Zweitklässlern insgesamt schneiden im Kompetenzbereich Mathematik [Lesen] …

Je weiter links Sie Ihr Kreuz machen, desto schlechter schneidet die Gruppe Ihrer Einschätzung nach ab, je weiter rechts Sie Ihr Kreuz machen, desto besser schneidet die Gruppe ab. Bitte in jeder Zeile ein Kästchen ankreuzen.

		sehr schlecht ab										sehr gut ab
		0	1	2	3	4	5	6	7	8	9	10
a)	… Mädchen	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
b)	... Jungen	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
c)	... Kinder aus niedrigen sozialen Schichten	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
d)	... Kinder aus mittleren sozialen Schichten	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
e)	... Kinder aus hohen sozialen Schichten	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
f)	… Kinder mit Migrations-hintergrund	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
g)	... Kinder mit türkischem Migrationshintergrund	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
h)	… Kinder mit russischem Migrationshintergrund	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
i)	... Kinder ohne Migrationshintergrund	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Figure C.1: German original of the first version of the instrument to measure teachers' stereotypes in the NEPS. Figure adopted from Wenz et al. (2016)

In der NEPS-Studie „Bildungsverläufe in Deutschland“ werden die Kompetenzen von Kindern in der zweiten Klasse in unterschiedlichen Bereichen erfasst.

Was denken Sie, wie Zweitklässler aus den folgenden Gruppen im Kompetenzbereich Mathematik [Lesen] im Vergleich zum Durchschnitt abschneiden werden?

Je weiter links Sie Ihr Kreuz machen, desto schlechter schneidet die Gruppe Ihrer Einschätzung nach ab, je weiter rechts Sie Ihr Kreuz machen, desto besser schneidet die Gruppe ab. Bitte in jeder Zeile ein Kästchen ankreuzen.

	sehr schlecht 0					5					sehr gut 10
a) Mädchen	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
b) Jungen	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Und wie werden die folgenden Gruppen im Vergleich zum Durchschnitt abschneiden?

	sehr schlecht 0					5					sehr gut 10
c) Kinder aus niedrigen sozialen Schichten	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
d) Kinder aus mittleren sozialen Schichten	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
e) Kinder aus hohen sozialen Schichten	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Und wie werden die folgenden Gruppen im Vergleich zum Durchschnitt abschneiden?

	sehr schlecht 0					5					sehr gut 10
f) Kinder mit Migrationshintergrund	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
g) Kinder mit türkischem Migrationshintergrund	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
h) Kinder mit russischem Migrationshintergrund	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
i) Kinder ohne Migrationshintergrund	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Figure C.2: German original of the second version of the instrument to measure teachers' stereotypes in the NEPS. Figure adopted from Wenz et al. (2016)

In der NEPS-Studie „Bildungsverläufe in Deutschland“ werden die Kompetenzen von Kindern in der zweiten Klasse in unterschiedlichen Bereichen erfasst.

Was denken Sie, welche Ergebnisse Zweitklässler aus folgenden Gruppen im Kompetenzbereich Mathematik [Lesen] im Vergleich zu Zweitklässlern in Deutschland insgesamt erzielen?

Je weiter links Sie Ihr Kreuz machen, desto schlechter werden die Ergebnisse der Gruppe Ihrer Einschätzung nach ausfallen, je weiter rechts Sie Ihr Kreuz machen, desto besser werden sie ausfallen. Bitte in jeder Zeile ein Kästchen ankreuzen.

	weit unter-durchschnittliche 0	1	2	3	4	5	6	7	8	9	weit über-durchschnittliche 10
a) Kinder aus niedrigen sozialen Schichten	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
b) Kinder aus mittleren sozialen Schichten	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
c) Kinder aus hohen sozialen Schichten	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Und welche Ergebnisse erzielen Zweitklässler aus folgenden Gruppen im Vergleich zu Zweitklässlern in Deutschland insgesamt?

	weit unter-durchschnittliche 0	1	2	3	4	5	6	7	8	9	weit über-durchschnittliche 10
d) Mädchen	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
e) Jungen	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Und welche Ergebnisse erzielen Zweitklässler aus folgenden Gruppen im Vergleich zu Zweitklässlern in Deutschland insgesamt?

	weit unter-durchschnittliche 0	1	2	3	4	5	6	7	8	9	weit über-durchschnittliche 10
f) Kinder mit Migrationshintergrund	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
g) Kinder mit türkischem Migrationshintergrund	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
h) Kinder mit russischem Migrationshintergrund	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
i) Kinder ohne Migrationshintergrund	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Figure C.3: German original of the final version of the instrument to measure teachers' stereotypes in the NEPS. Figure adopted from Wenz et al. (2016)

D Material Used in the Experiment

Vielen Dank, dass Sie sich zur Teilnahme an unserer Studie entschlossen haben!

Wir bitten Sie im Folgenden, einen kleinen Aufsatz (ca. 200 Wörter) zu bewerten und anschließend ein paar kurze Fragen zu Ihrem Beruf und Ihrer Person zu beantworten. Alle Angaben sind komplett anonym und werden ausschließlich für wissenschaftliche Zwecke verwendet.

Wenn Sie Informationen über die Ergebnisse unserer Studie erhalten oder an unserer Verlosung teilnehmen wollen, können Sie uns am Ende des Fragebogens Ihre E-Mail-Adresse hinterlassen. Die Adresse wird getrennt von Ihren anderen Angaben gespeichert und nicht an Dritte weitergegeben. Die Angabe ist freiwillig.

Die Teilnahme an der Studie wird nur wenige Minuten in Anspruch nehmen. Wir möchten Sie bitten, während dieser Zeit ein eventuelles Handy sowie Messenger- und Chatfunktionen auszuschalten. Herzlichen Dank!

Durch den Fragebogen bewegen Sie sich mit Hilfe der angegebenen Buttons. Ein Zurückblättern auf bereits bearbeitete Seiten ist nicht möglich.

Weiter

Figure D.1: First screen: Introductory screen with explanations of procedure.

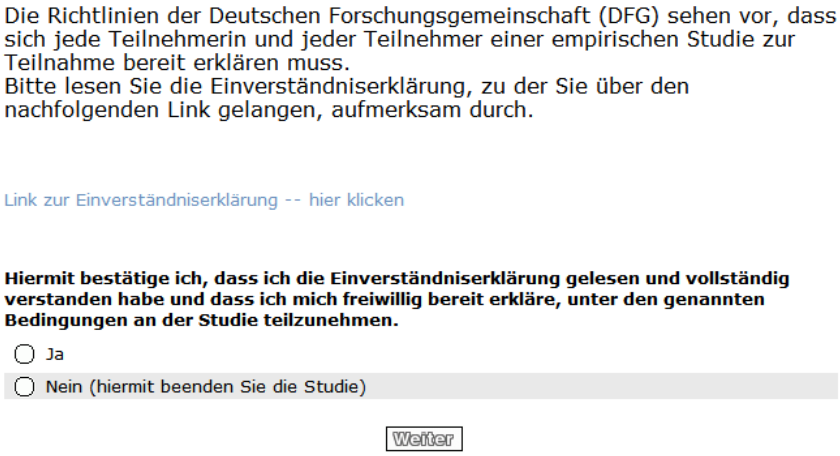
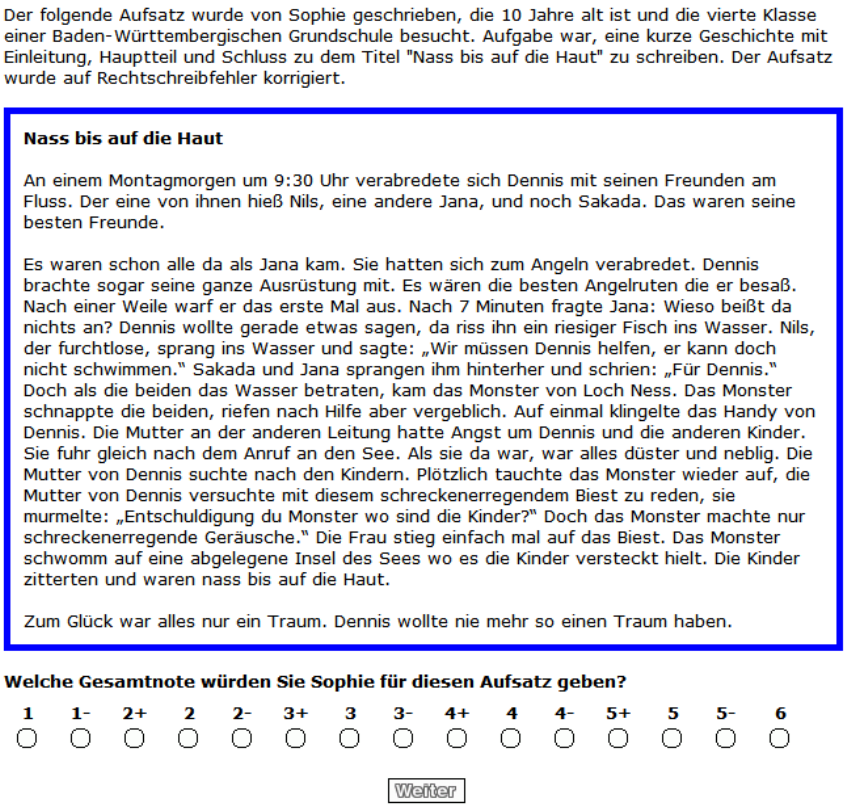


Figure D.2: Second screen: Consent form of the *Deutsche Forschungsgemeinschaft* (DFG).


Der folgende Aufsatz wurde von Sophie geschrieben, die 10 Jahre alt ist und die vierte Klasse einer Baden-Württembergischen Grundschule besucht. Aufgabe war, eine kurze Geschichte mit Einleitung, Hauptteil und Schluss zu dem Titel "Nass bis auf die Haut" zu schreiben. Der Aufsatz wurde auf Rechtschreibfehler korrigiert.

Nass bis auf die Haut

An einem Montagmorgen um 9:30 Uhr verabredete sich Dennis mit seinen Freunden am Fluss. Der eine von ihnen hieß Nils, eine andere Jana, und noch Sakada. Das waren seine besten Freunde.

Es waren schon alle da als Jana kam. Sie hatten sich zum Angeln verabredet. Dennis brachte sogar seine ganze Ausrüstung mit. Es wären die besten Angelruten die er besaß. Nach einer Weile warf er das erste Mal aus. Nach 7 Minuten fragte Jana: Wieso beißt da nichts an? Dennis wollte gerade etwas sagen, da riss ihn ein riesiger Fisch ins Wasser. Nils, der furchtlose, sprang ins Wasser und sagte: „Wir müssen Dennis helfen, er kann doch nicht schwimmen." Sakada und Jana sprangen ihm hinterher und schrien: „Für Dennis." Doch als die beiden das Wasser betraten, kam das Monster von Loch Ness. Das Monster schnappte die beiden, riefen nach Hilfe aber vergeblich. Auf einmal klingelte das Handy von Dennis. Die Mutter an der anderen Leitung hatte Angst um Dennis und die anderen Kinder. Sie fuhr gleich nach dem Anruf an den See. Als sie da war, war alles düster und neblig. Die Mutter von Dennis suchte nach den Kindern. Plötzlich tauchte das Monster wieder auf, die Mutter von Dennis versuchte mit diesem schreckenerregendem Biest zu reden, sie murmelte: „Entschuldigung du Monster wo sind die Kinder?" Doch das Monster machte nur schreckenerregende Geräusche." Die Frau stieg einfach mal auf das Biest. Das Monster schwomm auf eine abgelegene Insel des Sees wo es die Kinder versteckt hielt. Die Kinder zitterten und waren nass bis auf die Haut.

Zum Glück war alles nur ein Traum. Dennis wollte nie mehr so einen Traum haben.

Welche Gesamtnote würden Sie Sophie für diesen Aufsatz geben?

1	1-	2+	2	2-	3+	3	3-	4+	4	4-	5+	5	5-	6
○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Weiter

Figure D.3: Third screen: Text containing randomly allocated stimulus (here: Sophie) on top. Blue box containing one of the essays (here: good essay). Question on the bottom assesses overall grade for the essay.

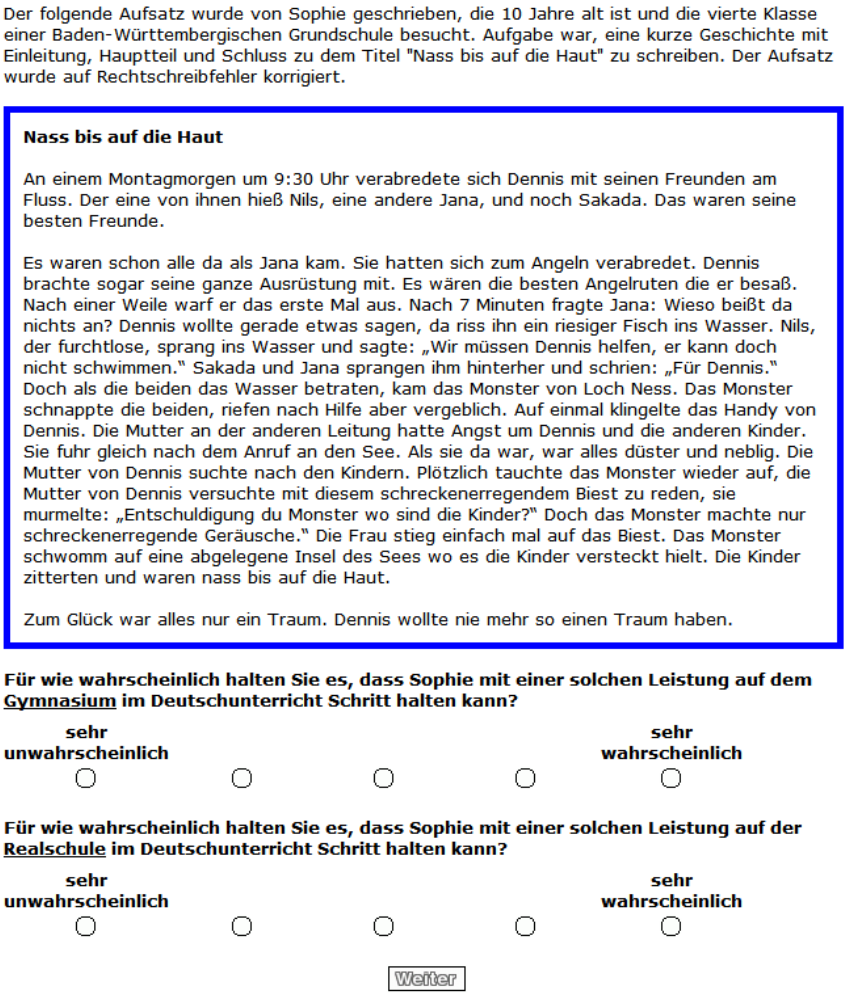
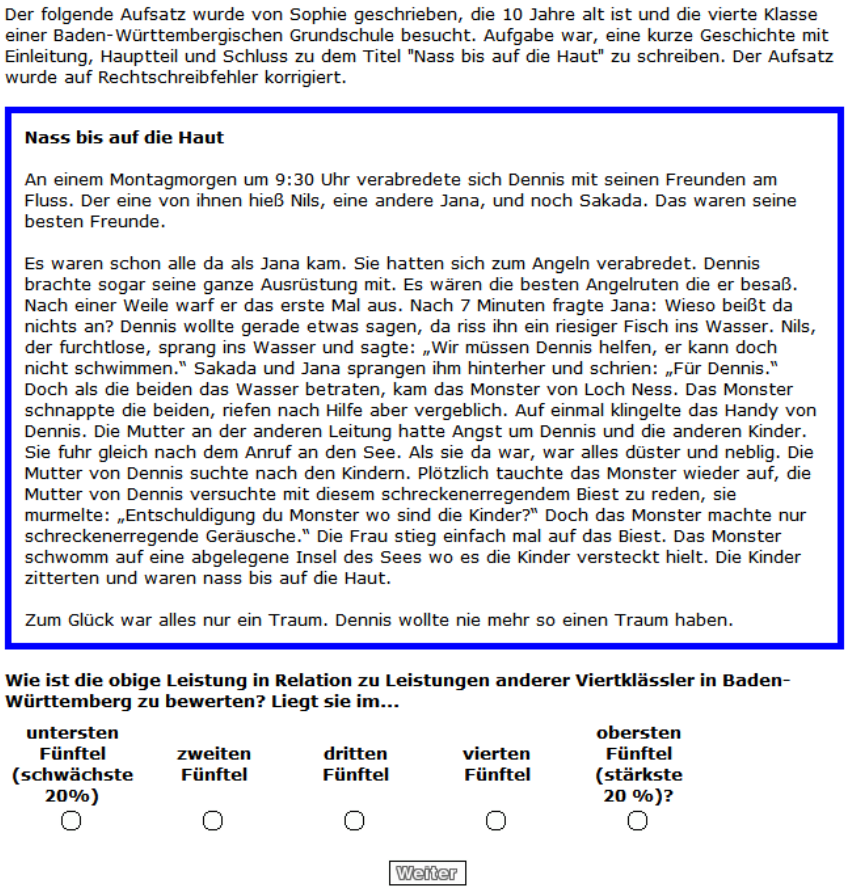


Figure D.4: Fourth screen: Text containing randomly allocated stimulus (here: Sophie) on top. Blue box containing one of the essays (here: good essay). Question on the bottom assesses overall grade for the essay.



Der folgende Aufsatz wurde von Sophie geschrieben, die 10 Jahre alt ist und die vierte Klasse einer Baden-Württembergischen Grundschule besucht. Aufgabe war, eine kurze Geschichte mit Einleitung, Hauptteil und Schluss zu dem Titel "Nass bis auf die Haut" zu schreiben. Der Aufsatz wurde auf Rechtschreibfehler korrigiert.

Nass bis auf die Haut

An einem Montagmorgen um 9:30 Uhr verabredete sich Dennis mit seinen Freunden am Fluss. Der eine von ihnen hieß Nils, eine andere Jana, und noch Sakada. Das waren seine besten Freunde.

Es waren schon alle da als Jana kam. Sie hatten sich zum Angeln verabredet. Dennis brachte sogar seine ganze Ausrüstung mit. Es wären die besten Angelruten die er besaß. Nach einer Weile warf er das erste Mal aus. Nach 7 Minuten fragte Jana: Wieso beißt da nichts an? Dennis wollte gerade etwas sagen, da riss ihn ein riesiger Fisch ins Wasser. Nils, der furchtlose, sprang ins Wasser und sagte: „Wir müssen Dennis helfen, er kann doch nicht schwimmen." Sakada und Jana sprangen ihm hinterher und schrien: „Für Dennis." Doch als die beiden das Wasser betraten, kam das Monster von Loch Ness. Das Monster schnappte die beiden, riefen nach Hilfe aber vergeblich. Auf einmal klingelte das Handy von Dennis. Die Mutter an der anderen Leitung hatte Angst um Dennis und die anderen Kinder. Sie fuhr gleich nach dem Anruf an den See. Als sie da war, war alles düster und neblig. Die Mutter von Dennis suchte nach den Kindern. Plötzlich tauchte das Monster wieder auf, die Mutter von Dennis versuchte mit diesem schreckenerregendem Biest zu reden, sie murmelte: „Entschuldigung du Monster wo sind die Kinder?" Doch das Monster machte nur schreckenerregende Geräusche." Die Frau stieg einfach mal auf das Biest. Das Monster schwomm auf eine abgelegene Insel des Sees wo es die Kinder versteckt hielt. Die Kinder zitterten und waren nass bis auf die Haut.

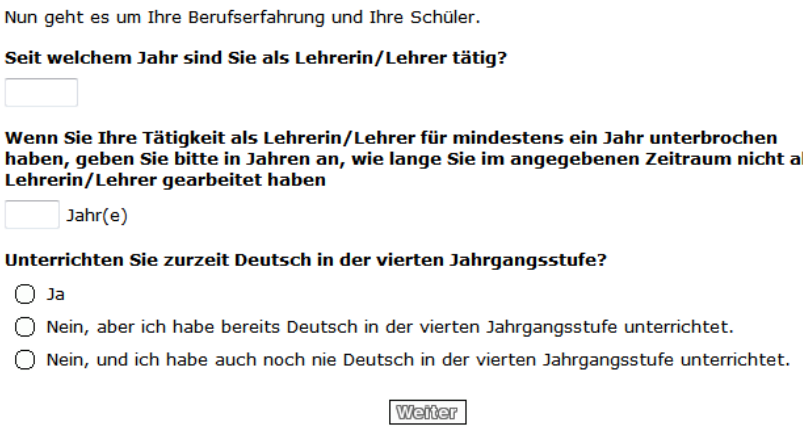
Zum Glück war alles nur ein Traum. Dennis wollte nie mehr so einen Traum haben.

Wie ist die obige Leistung in Relation zu Leistungen anderer Viertklässler in Baden-Württemberg zu bewerten? Liegt sie im...

untersten Fünftel (schwächste 20%)	zweiten Fünftel	dritten Fünftel	vierten Fünftel	obersten Fünftel (stärkste 20 %)?
○	○	○	○	○

Weiter

Figure D.5: Fifth screen: Text containing randomly allocated stimulus (here: Sophie) on top. Blue box containing one of the essays (here: good essay). Question on the bottom assesses essay relative to other fourth graders in Baden-Württemberg.


Nun geht es um Ihre Berufserfahrung und Ihre Schüler.

Seit welchem Jahr sind Sie als Lehrerin/Lehrer tätig?

Wenn Sie Ihre Tätigkeit als Lehrerin/Lehrer für mindestens ein Jahr unterbrochen haben, geben Sie bitte in Jahren an, wie lange Sie im angegebenen Zeitraum nicht als Lehrerin/Lehrer gearbeitet haben

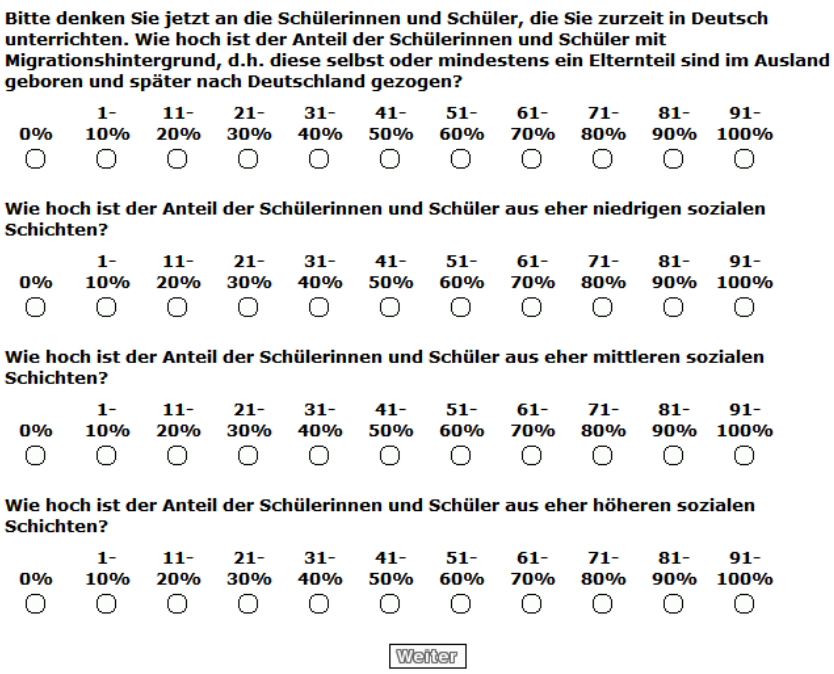
Jahr(e)

Unterrichten Sie zurzeit Deutsch in der vierten Jahrgangsstufe?

- ◯ Ja
- ◯ Nein, aber ich habe bereits Deutsch in der vierten Jahrgangsstufe unterrichtet.
- ◯ Nein, und ich habe auch noch nie Deutsch in der vierten Jahrgangsstufe unterrichtet.

Weiter

Figure D.6: Sixth screen: Questions on work experience as teacher, longer breaks from work, and experience in teaching German to fourth graders.


Bitte denken Sie jetzt an die Schülerinnen und Schüler, die Sie zurzeit in Deutsch unterrichten. Wie hoch ist der Anteil der Schülerinnen und Schüler mit Migrationshintergrund, d.h. diese selbst oder mindestens ein Elternteil sind im Ausland geboren und später nach Deutschland gezogen?

0%	1-10%	11-20%	21-30%	31-40%	41-50%	51-60%	61-70%	71-80%	81-90%	91-100%
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

Wie hoch ist der Anteil der Schülerinnen und Schüler aus eher niedrigen sozialen Schichten?

0%	1-10%	11-20%	21-30%	31-40%	41-50%	51-60%	61-70%	71-80%	81-90%	91-100%
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

Wie hoch ist der Anteil der Schülerinnen und Schüler aus eher mittleren sozialen Schichten?

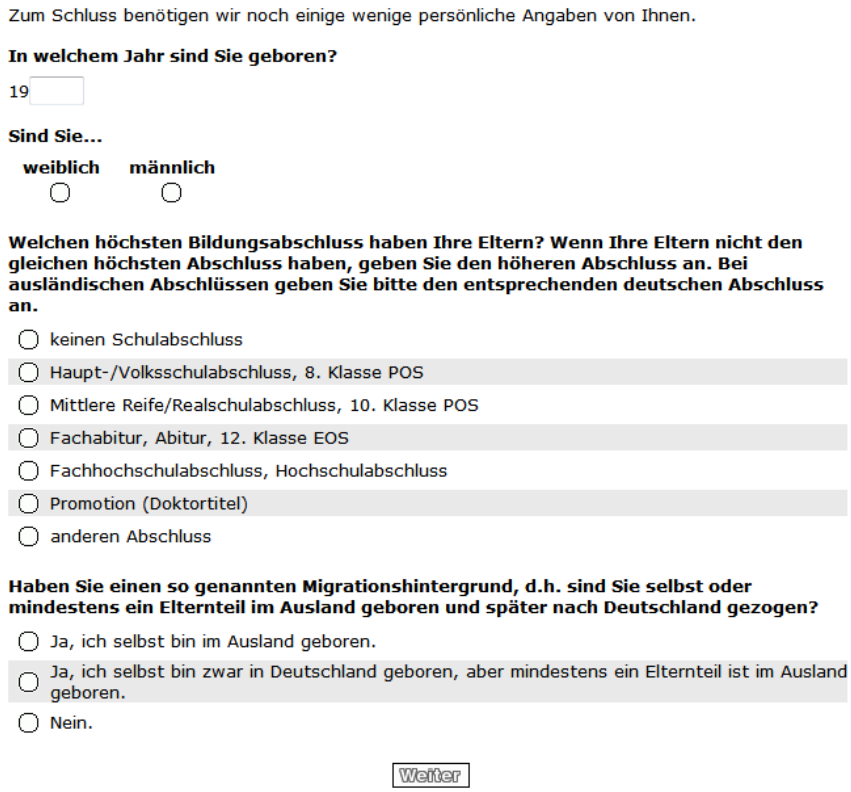
0%	1-10%	11-20%	21-30%	31-40%	41-50%	51-60%	61-70%	71-80%	81-90%	91-100%
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

Wie hoch ist der Anteil der Schülerinnen und Schüler aus eher höheren sozialen Schichten?

0%	1-10%	11-20%	21-30%	31-40%	41-50%	51-60%	61-70%	71-80%	81-90%	91-100%
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯

Weiter

Figure D.7: Seventh screen: Questions on proportion of students with immigrant background, lower class background, middle class background, and higher class background in classes taught by the teacher.


Zum Schluss benötigen wir noch einige wenige persönliche Angaben von Ihnen.

In welchem Jahr sind Sie geboren?

19

Sind Sie...

weiblich **männlich**

Welchen höchsten Bildungsabschluss haben Ihre Eltern? Wenn Ihre Eltern nicht den gleichen höchsten Abschluss haben, geben Sie den höheren Abschluss an. Bei ausländischen Abschlüssen geben Sie bitte den entsprechenden deutschen Abschluss an.

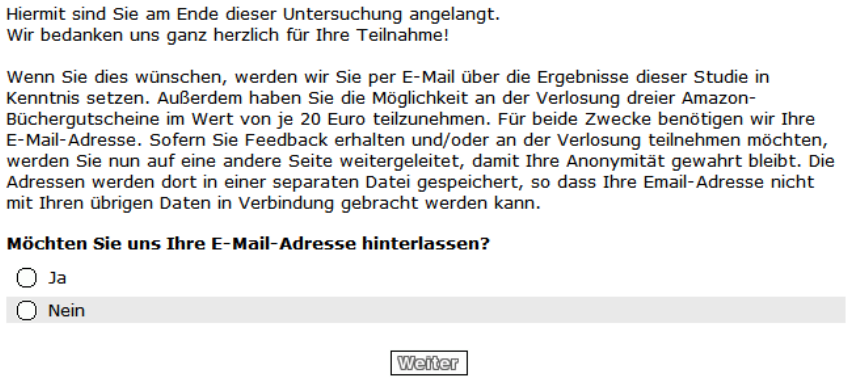
- keinen Schulabschluss
- Haupt-/Volksschulabschluss, 8. Klasse POS
- Mittlere Reife/Realschulabschluss, 10. Klasse POS
- Fachabitur, Abitur, 12. Klasse EOS
- Fachhochschulabschluss, Hochschulabschluss
- Promotion (Doktortitel)
- anderen Abschluss

Haben Sie einen so genannten Migrationshintergrund, d.h. sind Sie selbst oder mindestens ein Elternteil im Ausland geboren und später nach Deutschland gezogen?

- Ja, ich selbst bin im Ausland geboren.
- Ja, ich selbst bin zwar in Deutschland geboren, aber mindestens ein Elternteil ist im Ausland geboren.
- Nein.

Weiter

Figure D.8: Eighth screen: Questions on the demographics of the teacher: year of birth, sex/gender, highest education of parents, immigrant background.


Hiermit sind Sie am Ende dieser Untersuchung angelangt.
Wir bedanken uns ganz herzlich für Ihre Teilnahme!

Wenn Sie dies wünschen, werden wir Sie per E-Mail über die Ergebnisse dieser Studie in Kenntnis setzen. Außerdem haben Sie die Möglichkeit an der Verlosung dreier Amazon-Büchergutscheine im Wert von je 20 Euro teilzunehmen. Für beide Zwecke benötigen wir Ihre E-Mail-Adresse. Sofern Sie Feedback erhalten und/oder an der Verlosung teilnehmen möchten, werden Sie nun auf eine andere Seite weitergeleitet, damit Ihre Anonymität gewahrt bleibt. Die Adressen werden dort in einer separaten Datei gespeichert, so dass Ihre Email-Adresse nicht mit Ihren übrigen Daten in Verbindung gebracht werden kann.

Möchten Sie uns Ihre E-Mail-Adresse hinterlassen?

- Ja
- Nein

Weiter

Figure D.9: Ninth screen: Participants are thanked for participating in the study and asked whether they would like to leave their e-mail address to receive feedback about the study's results and/or take part in the lottery.

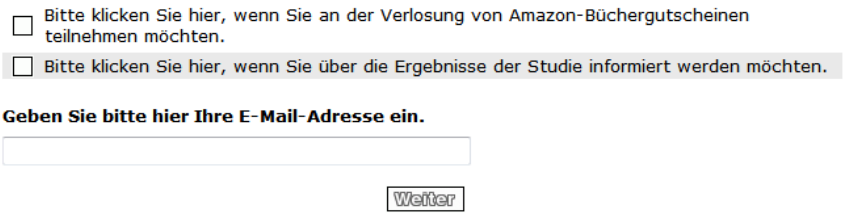


Figure D.10: Tenth screen: Participants may choose to receive feedback about the study's results and/or to take part in the lottery and share their e-mail address.

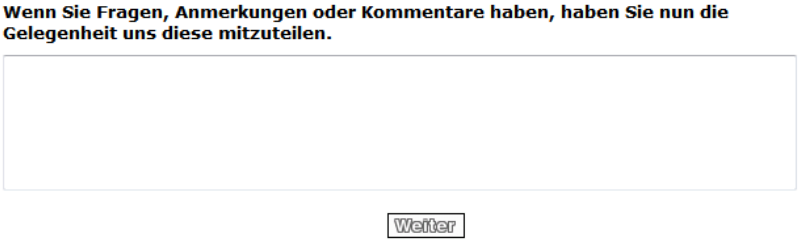


Figure D.11: Eleventh screen: Participants may share questions, remarks, or comments in an open-ended format.

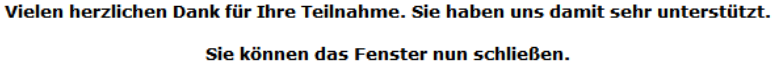


Figure D.12: Twelfth and final screen: Participants are thanked again and encouraged to close the window.

References

Abrams, D., & Hogg, M. A. (1988). Comments on the motivational status of self-esteem in social identity and intergroup discrimination. *European Journal of Social Psychology, 18*(4), 317–334. https://doi.org/10.1002/ejsp.2420180403/full

Agassi, J. (1975). Institutional individualism. *The British Journal of Sociology, 26*(2), 144–155. https://doi.org/10.2307/589585

Aigner, D., & Cain, G. (1977). Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review, 30*(2), 175–187.

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological Methods & Research, 28*(2), 186–208.

Allport, G. W. (1954). *The nature of prejudice*. Malden, MA, Addison-Wesley.

Altemeyer, B. (1981). *Right-wing authoritarianism*. University of Manitoba Press.

Altonji, J. G., & Blank, R. M. (1999). Race and gender in the labor market. In Orley C. Ashenfelter and David Card (Ed.), *Handbook of labor economics* (pp. 3143–3259). Elsevier. https://doi.org/10.1016/S1573-4463(99)30039-0

American Psychological Association. (2006). APA resolution on prejudice, stereotypes, and discrimination. Retrieved May 27, 2017, from https://www.apa.org/about/policy/prejudice.pdf

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, Princeton University Press.

Antonovsky, A. (1960). The social meaning of discrimination. *Phylon (1960-), 21*(1), 81–95. https://doi.org/10.2307/273741

Arnold, K.-H., Bos, W., Richert, P., & Stubbe, T. (2007). Schullaufbahnpräferenzen am Ende der vierten Klassenstufe. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, K. Schwippert, & R. Valtin (Eds.), *IGLU 2006: Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (pp. 271–298). Münster, Waxmann.

Arrow, K. J. (1973). The theory of discrimination. In O. Ashenfelter & A. Rees (Eds.), *Discrimination in labor markets* (pp. 3–33). Princeton, NJ, Princeton University Press.

Arrow, K. J. (1998). What has economics to say about racial discrimination? *Journal of Economic Perspectives, 12*, 91–100.

Artelt, C., & Rausch, T. (2014). Accuracy of teacher judgments. In S. Krolak-Schwerdt, S. Glock, & M. Böhmer (Eds.), *Teachers' professional development* (pp. 27–43). SensePublishers. https://doi.org/10.1007/978-94-6209-536-6_3

Ashenfelter, O., & Ham, J. (1979). Education, unemployment, and earnings. *Journal of Political Economy*, *87*(5), S99–S116. https://doi.org/10.1086/260824

Ashmore, R. D., & Del Boca, F. K. (1979). Sex stereotypes and implicit personality theory: Toward a cognitive–social psychological conceptualization. *Sex Roles*, *5*(2), 219–248. https://doi.org/10.1007/BF00287932

Ashmore, R. D., & Del Boca, F. K. (1981). Conceptual approaches to stereotypes and stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 1–35). Hillsdale, NJ, Erlbaum.

Ashmore, R. D., & Longo, L. C. (1995). Accuracy of stereotypes: What research on physical attractiveness can teach us. In Y.-T. Lee, L. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy. Toward appreciating group differences* (pp. 60–86). Washington, DC, American Psychological Association (APA).

Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2014). The rules of implicit evaluation by race, religion, and age. *Psychological Science*, 1804–1815. https://doi.org/10.1177/0956797614543801

Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science*, *7*(3), 136–141. https://doi.org/10.1111/j.1467-9280.1996.tb00346.x

Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological Bulletin*, *91*(1), 3–26. https://doi.org/10.1037/0033-2909.91.1.3

Baumeister, R. F., Tice, D. M., & Hutton, D. G. (1989). Self-presentational motivations and personality differences in self-esteem. *Journal of Personality*, *57*(3), 547–579. https://doi.org/10.1111/j.1467-6494.1989.tb02384.x/full

Becker, G. (1971). *The economics of discrimination* (2nd ed.). Chicago, IL, University of Chicago Press. (Original work published 1957)

Bell, M. (2008). The implementation of European anti-discrimination directives: Converging towards a common model? *The Political Quarterly*, *79*(1), 36–44. https://doi.org/10.1111/j.1467-923X.2008.00900.x

Below, S. v. (2007). What are the chances of young Turks and Italians for equal education and employment in Germany? The role of objective and subjective indicators. *Social Indicators Research*, *82*(2), 209–231. https://doi.org/10.1007/s11205-006-9038-6

Berard, T. J. (2008). The neglected social psychology of institutional racism. *Sociology Compass*, *2*(2), 734–764. https://doi.org/10.1111/j.1751-9020.2007.00089.x/full

Bergh, L. v. d., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal*, *47*(2), 497–527. https://doi.org/10.3102/0002831209353594

Berk, R. A. (2004). *Regression analysis: A constructive critique*. Thousand Oaks, CA, Sage.

Bertrand, M., Chugh, D., & Mullainathan, S. (2005). Implicit discrimination. *The American Economic Review, 95*(2), 94–98. https://doi.org/10.2307/4132797

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review, 94*(4), 991–1013. https://doi.org/10.1257/0002828042002561

Biernat, M., & Crandall, C. S. (1999). Racial attitudes. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of political attitudes* (1st, pp. 297–411). San Diego, Academic Press.

Black, D. A. (1995). Discrimination in an equilibrium search model. *Journal of Labor Economics, 13*(2), 309–334. https://doi.org/10.1086/298376

Blalock, H. M. J. (1967). *Toward a theory of minority-group relations*. New York, London, Sydney, John Wiley & Sons.

Blank, R. M., Dabady, M., & Citro, C. F. (Eds.). (2004). *Measuring racial discrimination*. Washington, D.C., The National Academies Press.

Blau, P. M., & Duncan, O. D. (1967). *The American occupational structure*. New York, John Wiley & Sons.

Bless, H., Fiedler, K., & Strack, F. (2004). *Social cognition: How individuals construct social reality* (1st ed.). Psychology Press.

Bless, H., & Schwarz, N. (2010). Mental construal and the emergence of assimilation and contrast effects: The inclusion/exclusion model. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 319–373). Academic Press.

Blumer, H. (1958). Race prejudice as a sense of group position. *The Pacific Sociological Review, 1*(1), 3–7. https://doi.org/10.2307/1388607

Blundell, R., Dearden, L., Meghir, C., & Sianesi, B. (1999). Human capital investment: The returns from education and training to the individual, the firm and the economy. *Fiscal Studies, 20*(1), 1–23. https://doi.org/10.1111/j.1475-5890.1999.tb00001.x

Bobo, L. D. (1999). Prejudice as group position: Microfoundations of a sociological approach to racism and race relations. *Journal of Social Issues, 55*(3), 445–472. https://doi.org/10.1111/0022-4537.00127

Bobo, L. D., & Fox, C. (2003). Race, racism, and discrimination: Bridging problems, methods, and theory in social psychological research. *Social Psychology Quarterly, 66*(4), 319–332. https://doi.org/10.2307/1519832

Bobo, L. D., & Hutchings, V. L. (1996). Perceptions of racial group competition: Extending Blumer's theory of group position to a multiracial social context. *American Sociological Review, 61*(6), 951–972. https://doi.org/10.2307/2096302

Bodenhausen, G. V., & Lichtenstein, M. (1987). Social stereotypes and information-processing strategies: The impact of task complexity. *Journal of Personality, 52*(5), 871–880. https://doi.org/10.1037/0022-3514.52.5.871

Bogardus, E. S. (1925). Social distance and its origin. *Journal of Applied Sociology, 9*, 216–226.

Bogardus, E. S. (1933). A social distance scale. *Sociology and Social Research, 17*, 265–271.

Bogardus, E. S. (1958). Racial distance changes in the United States during the past 30 years. *Sociology & Social Research, 43*, 127–134.

Bohren, J. A., Haggag, K., Imas, A., & Pope, D. G. (2019). *Inaccurate statistical discrimination* (Working Paper No. 25935). National Bureau of Economic Research. https://doi.org/10.3386/w25935

Böltken, F. (2000). Soziale Distanz und raumliche Nähe – Einstellungen und Erfahrungen im alltäglichen Zusammenleben von Ausländern und Deutschen im Wohngebiet. In R. Alba, P. Schmidt, & M. Wasmer (Eds.), *Deutsche und Ausländer: Freunde, Fremde oder Feinde?* (pp. 147–194). Wiesbaden, VS Verlag für Sozialwissenschaften.

Bonefeld, M., & Dickhäuser, O. (2018). (biased) grading of students' performance: Students' names, performance level, and implicit attitudes. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.00481

Bonilla-Silva, E. (1997). Rethinking racism: Toward a structural interpretation. *American Sociological Review, 62*(3), 465–480. https://doi.org/10.2307/2657316

Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics, 131*(4), 1753–1794. https://doi.org/10.1093/qje/qjw029

Bos, W., Tarelli, I., Bremerich-Vos, A., & Schwippert, K. (Eds.). (2012). *IGLU 2011: Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster, Waxmann.

Bos, W., Wendt, H., Köller, O., & Selter, C. (Eds.). (2012). *TIMSS 2011: Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster u.a., Waxmann.

Boudon, R. (1986a). Individualism and holism in the social sciences. In P. Birnbaum & J. Leca (Eds.), *Individualism: Theories and methods*. Oxford, Clarendon.

Boudon, R. (1986b). *Theories of social change: A critical appraisal*. University of California Press.

Boudon, R. (1998). Limitations of rational choice theory. *American Journal of Sociology, 104*(3), 817–828. https://doi.org/10.1086/210087

Boudon, R. (2003). Beyond rational choice theory. *Annual Review of Sociology, 29*(1), 1–21. https://doi.org/10.1146/annurev.soc.29.010202.100213

Brand, J. E., & Xie, Y. (2010). Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education. *American Sociological Review, 75*(2), 273–302. https://doi.org/10.1177/0003122410363567

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics, 46*(4), 1171–1178.

Brauer, M. (2001). Intergroup perception in the social context: The effects of social status and group membership on perceived out-group homogeneity and ethnocentrism. *Journal of Experimental Social Psychology, 37*(1), 15–31.

Breen, R., & Jonsson, J. O. (2005). Inequality of opportunity in comparative perspective: Recent research on educational attainment and social mobility. *Annual Review of Sociology*, 223–243. https://doi.org/10.2307/29737718

Breen, R., Luijkx, R., Müller, W., & Pollak, R. (2009). Nonpersistent inequality in educational attainment: Evidence from eight European countries. *American Journal of Sociology, 114*(5), 1475–1521. https://doi.org/10.1086/595951

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues, 55*(3), 429–444. https://doi.org/10.1111/0022-4537.00126/full

Brigham, J. C. (1971). Ethnic stereotypes. *Psychological Bulletin, 76*(1), 15–38. https://doi.org/10.1037/h0031446

Brown, R. (2010). *Prejudice: Its social psychology* (2nd ed.). Malden, MA, Wiley-Blackwell.

Brown, R., & Wootton-Millward, L. (1993). Perceptions of group homogeneity during group formation and change. *Social Cognition, 11*(1), 126–149. https://doi.org/10.1521/soco.1993.11.1.126

Brunello, G., Fabbri, D., & Fort, M. (2013). The causal effect of education on body mass: Evidence from Europe. *Journal of Labor Economics, 31*(1), 195–223. https://doi.org/10.1086/667236

Brunello, G., Fort, M., Schneeweis, N., & Winter-Ebmer, R. (2016). The causal effect of education on health: What is the role of health behaviors? *Health Economics, 25*(3), 314–336. https://doi.org/10.1002/hec.3141

Büchel, F., & Frick, J. R. (2004). Immigrants in the UK and in West Germany–relative income position, income portfolio, and redistribution effects. *Journal of Population Economics, 17*(3), 553–581.

Buchholz, S., & Schier, A. (2015). New game, new chance? Social inequalities and upgrading secondary school qualifications in West Germany. *European Sociological Review*, 603–615. https://doi.org/10.1093/esr/jcv062

Cain, G. G. (1986). The economic analysis of labor market discrimination: A survey. In Orley C. Ashenfelter and Richard Layard (Ed.), *Handbook of labor economics* (pp. 693–785). Elsevier.

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition. A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review*, *16*(4), 330–350. https://doi.org/10.1177/1088868312440047

Campbell, T. (2015). Stereotyped at seven? Biases in teacher judgement of pupils' ability and attainment. *Journal of Social Policy*, *44*(3), 517–547. https://doi.org/10.1017/S0047279415000227

Caporael, L. R. (1997). The evolution of truly social cognition: The core configurations model. *Personality and Social Psychology Review*, *1*(4), 276–298. https://doi.org/10.1207/s15327957pspr0104_1

Card, D. (1999). The causal effect of education on earnings. In D. Card & O. Ashenfelter (Eds.), *Handbook of labor economics* (pp. 1801–1863). Amsterdam, Elsevier. https://doi.org/10.1016/S1573-4463(99)03011-4

Carmichael, S., & Hamilton, C. V. (1967). *Black power: The politics of liberation in America*. New York, Random House.

Carneiro, P., Heckman, J. J., & Masterov, D. V. (2005). Labor market discrimination and racial differences in premarket factors. *The Journal of Law and Economics*, *48*(1), 1–39. https://doi.org/10.1086/426878

Carvacho, H., Zick, A., Haye, A., González, R., Manzi, J., Kocik, C., & Bertl, M. (2013). On the relation between social class and prejudice: The roles of education, income, and ideological attitudes. *European Journal of Social Psychology*, *43*(4), 272–285. https://doi.org/10.1002/ejsp.1961

Charles, K. K., & Guryan, J. (2011). Studying discrimination: Fundamental challenges and recent progress. *Annual Review of Economics*, *3*(1), 479–511. https://doi.org/10.1146/annurev.economics.102308.124448

Chiras, D., & Crea, D. (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *The American Economic Review*, *94*(1), 155–189. https://doi.org/10.1257/000282804322970751

Chopin, I., & Germaine, C. (2016). *A comparative analysis of non-discrimination law in Europe 2015. A comparative analysis of the implementation of EU non-discrimination law in the EU member states, the former Yugoslav Republic of Macedonia, Iceland,*

Liechtenstein, Montenegro, Norway, Serbia and Turkey. European Commission. Brussels. https://doi.org/10.2838/118488

Chun, W. Y., & Kruglanski, A. W. (2006). The role of task demands and processing resources in the use of base-rate and individuating information. *Journal of Personality and Social Psychology, 91*(2), 205–217. https://doi.org/10.1037/0022-3514.91.2.205

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Revised Edition). New York, Academic Press.

Colella, A., Hebl, M., & King, E. (2017). One hundred years of discrimination research in the Journal of Applied Psychology: A sobering synopsis. *Journal of Applied Psychology, 102*(3), 500–513. https://doi.org/10.1037/apl0000084

Coleman, J. S. (1986). Social theory, social research, and a theory of action. *American Journal of Sociology, 91*(6), 1309–1335. https://doi.org/10.1086/228423

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research, 12*(3), 229–238. https://doi.org/10.1023/A:1023254226592

Conti, G., Heckman, J., & Urzua, S. (2010). The education-health gradient. *The American Economic Review, 100*(2), 234–238.

Correll, J., Judd, C. M., Park, B., & Wittenbrink, B. (2010). Measuring prejudice, stereotypes, and discrimination. In J. F. Dovidio, M. Hewstone, P. Glick, & V. M. Esses (Eds.), *The SAGE handbook of prejudice, stereotyping, and discrimination* (pp. 45–62). Thousand Oaks, Sage.

Crandall, C. S., Bahns, A. J., Warner, R., & Schaller, M. (2011). Stereotypes as justifications of prejudice. *Personality and Social Psychology Bulletin, 37*(11), 1488–1498. https://doi.org/10.1177/0146167211411723

Crandall, C. S., & Stangor, C. (2005). Conformity and prejudice. In J. F. Dovidio, P. Glick, & L. A. Rudman (Eds.), *On the nature of prejudice* (pp. 293–309). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470773963.ch18/summary

Croizet, J.-C. (2008). The pernicious relationship between merit assessment and discrimination in education. In G. Adams, M. Biernat, & N. R. Branscombe (Eds.), *Commemorating Brown: The social psychology of racism and discrimination* (pp. 153–172). Washington, D.C., American Psychological Association.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology, 92*(4), 631–648. https://doi.org/10.1037/0022-3514.92.4.631

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*, *12*(2), 163–170. https://doi.org/10.1111/1467-9280.00328

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20–33. https://doi.org/10.1037/0022-3514.44.1.20

De Houwer, J., & Moors, A. (2007). How to define and examine the implicitness of implicit measures. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 59–102). New York, NY, The Guilford Press.

Dee, T. S. (2004a). Are there civic returns to education? *Journal of Public Economics*, *88*(9), 1697–1720. https://doi.org/10.1016/j.jpubeco.2003.11.002

Dee, T. S. (2004b). Teachers, race, and student achievement in a randomized experiment. *The Review of Economics and Statistics*, *86*(1), 195–210.

DeMeis, D. K., & Turner, R. R. (1978). Effects of students' race, physical attractiveness, and dialect on teachers' evaluations. *Contemporary Educational Psychology*, *3*(1), 77–86.

Deming, D. J., Yuchtman, N., Abulafi, A., Goldin, C., & Katz, L. F. (2016). The value of postsecondary credentials in the labor market: An experimental study. *American Economic Review*, *106*(3), 778–806. https://doi.org/10.1257/aer.20141757

Dempster, A. P. (1988). Employment discrimination and statistical science. *Statistical Science*, *3*(2), 149–161. https://doi.org/10.1214/ss/1177012894

DeSario, N. J. (2003). Reconceptualizing meritocracy: The decline of disparate impact discrimination law. *Harvard Civil Rights-civil Liberties Law Review*, *38*, 479–510.

Desimone, L. M., & Floch, K. C. L. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis*, *26*(1), 1–22. https://doi.org/10.3102/01623737026001001

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5–18.

Diefenbach, H. (2010). *Kinder und Jugendliche aus Migrantenfamilien im deutschen Bildungssystem* (3rd ed.). Wiesbaden, VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91042-0.pdf

Diehl, C., & Fick, P. (2016). Ethnische Diskriminierung im deutschen Bildungssystem. In C. Diehl, C. Hunkler, & C. Kristen (Eds.), *Ethnische Ungleichheiten im Bildungsverlauf* (pp. 243–286). Wiesbaden, Springer. https://doi.org/10.1007/978-3-658-04322-3_6

Diekman, A. B., Eagly, A. H., & Kulesa, P. (2002). Accuracy and bias in stereotypes about the social and political attitudes of women and men. *Journal of Experimental Social Psychology*, *38*(3), 268–282. https://doi.org/10.1006/jesp.2001.1511

Ditton, H. (2013). Wer geht auf die Hauptschule? Primäre und sekundäre Effekte der sozialen Herkunft beim Übergang nach der Grundschule. *Zeitschrift für Erziehungswissenschaft*, *16*(4), 731–749. https://doi.org/10.1007/s11618-013-0440-y

Ditton, H., & Aulinger, J. (2011). Schuleffekte und institutionelle Diskriminierung – eine kritische Auseinandersetzung mit Mythen und Legenden in der Schulforschung. In R. Becker (Ed.), *Integration durch Bildung* (pp. 95–119). Wiesbaden, VS Verlag für Sozialwissenschaften.

Ditton, H., Krüsken, J., & Schauenberg, M. (2005). Bildungsungleichheit – der Beitrag von Familie und Schule. *Zeitschrift für Erziehungswissenschaft*, *8*(2), 285–304. https://doi.org/10.1007/s11618-005-0138-x

Dolan, P., Peasgood, T., & White, M. (2008). Do we really know what makes us happy? a review of the economic literature on the factors associated with subjective well-being. *Journal of Economic Psychology*, *29*(1), 94–122. https://doi.org/10.1016/j.joep.2007.09.001

Dovidio, J. F., Brigham, J. C., Johnson, B. T., & Gaertner, S. L. (1996). Stereotyping, prejudice, and discrimination: Another look. In C. N. Macrae, C. Stangor, & M. Hewstone (Eds.), *Stereotypes and stereotyping*. New York, London, The Guilford Press.

Dovidio, J. F., & Gaertner, S. L. (2008). New directions in aversive racism research: Persistence and pervasiveness. In C. Willis-Esqueda (Ed.), *Motivational aspects of prejudice and racism* (pp. 43–67). Springer New York. https://doi.org/10.1007/978-0-387-73233-6_3

Dovidio, J. F., Gaertner, S. L., & Pearson, A. R. (2017). Aversive racism and contemporary bias. In *The cambridge handbook of the psychology of prejudice* (pp. 267–294). New York, NY, US, Cambridge University Press. https://doi.org/10.1017/9781316161579.012

Dovidio, J. F., Hewstone, M., Glick, P., & Esses, V. M. (2010). Prejudice, stereotyping, and discrimination: Theoretical and empirical overview. In J. F. Dovidio, M. Hewstone, P. Glick, & V. M. Esses (Eds.), *The SAGE handbook of prejudice, stereotyping, and discrimination* (pp. 45–62). Thousand Oaks, Sage.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*(1), 62.

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology*, *33*(5), 510–540. https://doi.org/10.1006/jesp.1997.1331

Driessen, G., Sleegers, P., & Smit, F. (2008). The transition from primary to secondary education: Meritocracy and ethnicity. *European Sociological Review*, *24*(4), 527–542. https://doi.org/10.1093/esr/jcn018

Dustmann, C., Frattini, T., & Lanzara, G. (2012). Educational achievement of second-generation immigrants: An international comparison. *Economic Policy*, *27*(69), 143–185. https://doi.org/10.1111/j.1468-0327.2011.00275.x

Ehrlich, H. J. (1973). *The social psychology of prejudice: A systematic theoretical review and propositional inventory of the American Social Psychological Study of Prejudice.* New York, John Wiley & Sons.

Elster, J. (1982). Marxism, functionalism, and game theory: The case for methodological individualism. *Theory and Society*, *11*(4), 453–482.

Elster, J. (1989). *Nuts and bolts for the social sciences* (1st ed.). Cambridge; New York, Cambridge University Press.

Elwert, F. (2013). Graphical causal models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 245–273). Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3_13

England, P., & Lewin, P. (1989). Economic and sociological views of discrimination in labor markets: Persistence or demise? *Sociological Spectrum*, *9*(3), 239–257.

Erikson, R., & Jonsson, J. O. (Eds.). (1996). *Can education be equalized? The Swedish case in comparative perspective*. Westview Press.

Esser, H. (1999). *Soziologie: Spezielle Grundlagen. Band 1: Situationslogik und Handeln*. Frankfurt; New York, Campus Verlag.

Esser, H. (2001). *Soziologie: Spezielle Grundlagen. Band 6: Sinn und Kultur.* Frankfurt; New York, Campus.

Esser, H. (2009). Rationality and commitment: The model of frame selection and the explanation of normative action. In M. Cherkaoui & P. Hamilton (Eds.), *Raymond Boudon: A life in sociology. Part two: Toward a general theory of rationality* (pp. 207–230). Oxford, Bardwell.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. *Advances in Experimental Social Psychology*, *23*, 75–109.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide

pipeline? *Journal of Personality and Social Psychology, 69*(6), 1013–1027. https://doi.org/10.1037/0022-3514.69.6.1013

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54*, 297–327. https://doi.org/10.1146/annurev.psych.54.101601.145225

Fazio, R. H., & Petty, R. E. (2008). *Attitudes: Their structure, function, and consequences*. New York, Psychology Press.

Feagin, J. R. (1977). Indirect institutionalized discrimination: A typological and policy analysis. *American Politics Quarterly, 5*(2), 177–200.

Feagin, J. R. (2006). *Systematic racism* (1st ed.). New York, Taylor & Francis.

Feagin, J. R., & Bennefield, Z. (2014). Systemic racism and U.S. health care. *Social Science & Medicine, 103*, 7–14. https://doi.org/10.1016/j.socscimed.2013.09.006

Feagin, J. R., & Booher Feagin, C. (1986). *Discrimination American style: Institutional racism and sexism* (2nd ed.). Malabar, FL, Krieger Publishing Company.

Feagin, J., & Eckberg, D. (1980). Discrimination: Motivation, action, effects, and context. *Annual Review of Sociology, 6*(1), 1–20.

Fein, S., & Spencer, S. (1997). Prejudice as self-image maintenance: Affirming the self through derogating others. *Journal of Personality and Social Psychology, 73*(1), 31–44.

Feldman, R. S., & Orchowsky, S. (1979). Race and performance of student as determinants of teacher nonverbal behavior. *Contemporary Educational Psychology, 4*(4), 324–333.

Fienberg, S. E., & Haviland, A. M. (2003). Discussion of 'statistics and causal inference: A review' (Pearl 2003). *Test, 12*(2), 319–327. https://doi.org/10.1007/BF02595718

Figlio, D. N. (2005). *Names, expectations and the black-white test score gap* (Working Paper No. 11195). National Bureau of Economic Research.

Fishbein, H. D. (2002, July 3). *Peer prejudice and discrimination: The origins of prejudice* (2nd). Mahwah, N.J, Lawrence Erlbaum Associates.

Fiske, S. T. (1993a). Controlling other people: The impact of power on stereotyping. *American Psychologist, 48*(6), 621–628. https://doi.org/10.1037/0003-066X.48.6.621

Fiske, S. T. (1993b). Social cognition and social perception. *Annual Review of Psychology, 44*, 155–194. https://doi.org/10.1146/annurev.ps.44.020193.001103

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed.). New York, McGraw-Hill.

Fiske, S. T. (2000). Stereotyping, prejudice, and discrimination at the seam between the centuries: Evolution, culture, mind, and brain. *European Journal of Social Psychology, 30*(3), 299–322.

Fiske, S. T., Lin, M., & Neuberg, S. L. (1999). The continuum model. Ten years later. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology.* (pp. 231–254). New York, Guilford.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology, 23*, 1–74. https://doi.org/10.1016/S0065-2601(08)60317-2

Fix, M., Galster, G. C., & Struyk, R. J. (1993). An overview of auditing for discrimination. In M. Fix & R. J. Struyk (Eds.), *Clear and convincing evidence: Measurement of discrimination in America* (pp. 1–67). Washington, D.C., The Urban Institute Press.

Fleischmann, F., Kristen, C., Heath, A. F., Brinbaum, Y., Deboosere, P., Granato, N., Jonsson, J. O., Kilpi-Jakonen, E., Lorenz, G., Lutz, A. C., Mos, D., Mutarrak, R., Phalet, K., Rothon, C., Rudolphi, F., & Werfhorst, H. G. v. d. (2014). Gender inequalities in the education of the second generation in western countries. *Sociology of Education, 87*(3), 143–170. https://doi.org/10.1177/0038040714537836

Fredman, S. (2012). *Comparative study of anti-discrimination and equality laws of the US, Canada, South Africa and India*. European Commission. Brussels.

Freedman, D. A. (2004). Graphical models for causation, and the identification problem. *Evaluation Review, 28*(4), 267–293.

Fryer Jr, R. G., & Levitt, S. D. (2004). The causes and consequences of distinctively black names. *The Quarterly Journal of Economics, 119*(3), 767–805.

Gaddis, S. M. (2015). Discrimination in the credential society: An audit study of race and college selectivity in the labor market. *Social Forces, 93*(4), 1451–1479. https://doi.org/10.1093/sf/sou111

Gaddis, S. M. (2017a). How black are lakisha and jamal? racial perceptions from names used in correspondence audit studies. *Sociological Science, 4*, 469–489. https://doi.org/10.15195/v4.a19

Gaddis, S. M. (2017b). Racial/ethnic perceptions from hispanic names: Selecting names to test for discrimination. *Socius, 3*, 2378023117737193. https://doi.org/10.1177/2378023117737193

Gaertner, S. L., & Dovidio, J. F. (1986). The aversive form of racism. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 61–89). Orlando, FL, Academic Press.

Gaertner, S. L., Dovidio, J. F., Anastasio, P. A., Bachman, B. A., & Rust, M. C. (1993). The common ingroup identity model: Recategorization and the reduction of intergroup bias. *European Review of Social Psychology, 4*(1), 1–26. https://doi.org/10.1080/14792779343000004

Gangl, M. (2010). Causal inference in sociological research. *Annual Review of Sociology, 36*(1), 21–47. https://doi.org/10.1146/annurev.soc.012809.102702

Ganter, S. (2003). *Soziale Netzwerke und interethnische Distanz: Theoretische und empirische Analysen zum Verhältnis von Deutschen und Ausländern*. Springer-Verlag.

Gardner, R. (1973). Ethnic stereotypes: The traditional approach, a new look. *Canadian Psychologist, 14*(2), 133–148.

Gawronski, B. (2009). Ten frequently asked questions about implicit measures and their frequently supposed, but not entirely correct answers. *Canadian Psychology/psychologie Canadienne, 50*(3), 141–150. https://doi.org/10.1037/a0013848

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692

Gawronski, B., & Creighton, L. A. (2013). Dual process theories. In D. Carlston (Ed.), *The oxford handbook of social cognition* (pp. 282–312). New York, NY, Oxford University Press.

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us? Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science, 2*(2), 181–193. https://doi.org/10.1111/j.1745-6916.2007.00036.x

Geis, A. (2011, March). *Handbuch für die Berufsvercodung*. Tech. rep. Mannheim, GESIS – Leibniz-Institute for the Social Sciences.

General Act on Equal Treatment (2006, August 18). Retrieved May 26, 2016, from http://www.gesetze-im-internet.de/agg/index.html

Glick, P., & Fiske, S. T. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology, 70*(3), 491–512. https://doi.org/10.1037/0022-3514.70.3.491

Glick, P., & Fiske, S. T. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American Psychologist, 56*(2), 109–118. https://doi.org/10.1037/0003-066X.56.2.109

Glock, S., & Karbach, J. (2015). Preservice teachers' implicit attitudes toward racial minority students: Evidence from three implicit measures. *Studies in Educational Evaluation, 45*, 55–61. https://doi.org/10.1016/j.stueduc.2015.03.006

Glock, S., & Klapproth, F. (2017). Bad boys, good girls? Implicit and explicit attitudes toward ethnic minority students among elementary and secondary school teachers. *Studies in Educational Evaluation, 53*, 77–86. https://doi.org/10.1016/j.stueduc.2017.04.002

Glock, S., Krolak-Schwerdt, S., & Pit-ten Cate, I. M. (2015). Are school placement recommendations accurate? the effect of students' ethnicity on teachers' judgments and recognition memory. *European Journal of Psychology of Education, 30*(2), 169–188. https://doi.org/10.1007/s10212-014-0237-2

Gomolla, M. (2010). Institutionelle Diskriminierung. Neue Zugänge zu einem alten Problem. In U. Hormel & A. Scherr (Eds.), *Diskriminierung* (pp. 61–93). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92394-9_4

Gomolla, M. (2016). Direkte und indirekte, institutionelle und strukturelle Diskriminierung. In A. Scherr, A. El-Mafaalani, & E. G. Yüksel (Eds.), *Handbuch Diskriminierung* (pp. 1–23). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-11119-9_9-1

Gomolla, M., & Radtke, F.-O. (2010). *Institutionelle Diskriminierung. Die Herstellung ethnischer Differenz in der Schule* (3rd ed.). Wiesbaden, VS Verlag für Sozialwissenschaften.

González, R., & Brown, R. (2003). Generalization of positive attitude as a function of subgroup and superordinate group identifications in intergroup contact. *European Journal of Social Psychology, 33*(2), 195–214. https://doi.org/10.1002/ejsp.140

Green, T. K. (2003). Discrimination in workplace dynamics: Toward a structural account of disparate treatment theory. *Harv. CR-CLL Rev., 38*, 91.

Greene, J. E. (1938). Analyses of racial differences within seven clinical categories of white and negro mental patients in the Georgia state hospital, 1923-32. *Social Forces, 17*(2), 201–211. https://doi.org/10.2307/2570925

Greenwald, A. G., Andrew, T., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17–41. https://doi.org/10.1037/a0015575

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review, 94*(4), 945–967. https://doi.org/10.2307/20439056

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of Personality and Social Psychology, 85*(2), 197.

Greenwald, A. G., & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist, 69*(7), 669. https://doi.org/10.1037/a0036056

Greiner, D. J., & Rubin, D. B. (2010). Causal effects of perceived immutable characteristics. *Review of Economics and Statistics, 93*(3), 775–785. https://doi.org/10.1162/REST_a_00110

Gresch, C. (2012). *Der Übergang in die Sekundarstufe I. Leistungsbeurteilung, Bildungsaspiration und rechtlicher Kontext bei Kindern mit Migrationshintergrund*. Wiesbaden, Springer VS.

Guryan, J., & Charles, K. K. (2013). Taste-based or statistical discrimination: The economics of discrimination returns to its roots. *The Economic Journal, 123*(572), F417–F432. https://doi.org/10.1111/ecoj.12080

Hachfeld, A., Hahn, A., Schroeder, S., Anders, Y., & Kunter, M. (2015). Should teachers be colorblind? How multicultural and egalitarian beliefs differentially relate to aspects of teachers' professional competence for teaching in diverse classrooms. *Teaching and Teacher Education, 48*, 44–55. https://doi.org/10.1016/j.tate.2015.02.001

Hachfeld, A., Hahn, A., Schroeder, S., Anders, Y., Stanat, P., & Kunter, M. (2011). Assessing teachers' multicultural and egalitarian beliefs: The teacher cultural beliefs scale. *Teaching and Teacher Education, 27*(6), 986–996. https://doi.org/10.1016/j.tate.2011.04.006

Hachfeld, A., Schroeder, S., Anders, Y., Hahn, A., & Kunter, M. (2012). Multikulturelle Überzeugungen. Herkunft oder Überzeugung? Welche Rolle spielen der Migrationshintergrund und multikulturelle Überzeugungen für das Unterrichten von Kindern mit Migrationshintergrund? *Zeitschrift für Pädagogische Psychologie, 26*(2), 101–120. https://doi.org/10.1024/1010-0652/a000064

Hagendoorn, L. (1995). Intergroup biases in multiple group systems: The perception of ethnic hierarchies. *European Review of Social Psychology, 6*(1), 199–228. https://doi.org/10.1080/14792779443000058

Hamermesh, D. S., & Biddle, J. E. (1993). *Beauty and the labor market* (Working Paper No. 4518). National Bureau of Economic Research. https://doi.org/10.3386/w4518

Hanna, R. N., & Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy, 4*(4), 146–68. https://doi.org/10.1257/pol.4.4.146

Harari, H., & McDavid, J. W. (1973). Name stereotypes and teachers' expectations. *Journal of Educational Psychology, 65*(2), 222–225.

Harmon, C., Oosterbeek, H., & Walker, I. (2003). The returns to education: Microeconomics. *Journal of Economic Surveys, 17*(2), 115–156. https://doi.org/10.1111/1467-6419.00191

Hasse, R., & Schmidt, L. (2012). Institutionelle Diskriminierung. In U. Bauer, U. H. Bittlingmayer, & A. Scherr (Eds.), *Handbuch Bildungs- und Erziehungssoziologie* (pp. 883–899). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-18944-4_52

Hausman, D. M. (1988). Ceteris paribus clauses and causality in economics. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1988*, 308–316.

Haynes, S. N., S, C., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Heath, A. F., Rothon, C., & Kilpi, E. (2008). The second generation in Western Europe: Education, unemployment, and occupational attainment. *Annual Review of Sociology, 34*(1), 211–235. https://doi.org/10.1146/annurev.soc.34.040507.134728

Heckman, J. J. (1998). Detecting discrimination. *The Journal of Economic Perspectives, 12*(2), 101–116. https://doi.org/10.2307/2646964

Heckman, J. J., & Siegelman, P. (1993). The urban institute audit studies: Their methods and findings. In M. Fix & R. J. Struyk (Eds.), *Clear and convincing evidence: Measurement of discrimination in America* (pp. 187–258). Washington, D.C., The Urban Institute Press.

Helbig, M., & Nikolai, R. (2015). *Die Unvergleichbaren. Der Wandel der Schulsysteme in den deutschen Bundesländern seit 1949*. Bad Heilbrunn, Verlag Julius Klinkhardt.

Henderson, J., & Chatfield, S. (2011). Who matches? Propensity scores and bias in the causal effects of education on participation. *The Journal of Politics, 73*(3), 646–658. https://doi.org/10.1017/S0022381611000351

Henry, P. J. (2010). Institutional bias. In J. F. Dovidio, M. Hewstone, P. Glick, & V. M. Esses (Eds.), *The SAGE handbook of prejudice, stereotyping, and discrimination* (pp. 426–440). Thousand Oaks, Sage.

Hill, P. B. (1984). Räumliche Nähe und soziale Distanz zu ethnischen Minderheiten. *Zeitschrift Für Soziologie, 13*(4), 363–370.

Hillmert, S., & Jacob, M. (2005). Institutionelle Strukturierung und inter-individuelle Variation. *Kölner Zeitschrift Für Soziologie Und Sozialpsychologie, 57*(3), 414–442. https://doi.org/10.1007/s11577-005-0183-8

Hillmert, S., & Jacob, M. (2010). Selections and social selectivity on the academic track: A life-course analysis of educational attainment in Germany. *Research in Social Stratification and Mobility, 28*(1), 59–76. https://doi.org/10.1016/j.rssm.2009.12.006

Hilton, J. L., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology, 47*, 237–271. https://doi.org/10.1146/annurev.psych.47.1.237

Hinnerich, B. T., Höglin, E., & Johannesson, M. (2011). Are boys discriminated in Swedish high schools? *Economics of Education Review, 30*(4), 682–690. https://doi.org/10.1016/j.econedurev.2011.02.007

Hinnerich, B. T., Höglin, E., & Johannesson, M. (2014). Discrimination against students with foreign backgrounds: Evidence from grading in Swedish public high schools. *Education Economics, 23*(6), 1–17. https://doi.org/10.1080/09645292.2014.899562

Hochweber, J. (2010). *Was erfassen Mathematiknoten? Korrelate von Mathematik-Zeugniszensuren auf Schüler- und Schulklassenebene in Primar- und Sekundarstufe.* Münster, Waxmann.

Hoenig, K., & Wenz, S. E. (2013). *Ethnic and social class discrimination in education. experimental evidence from germany* [ASA 2013 Annual Meeting]. New York, NY. ASA 2013 Annual Meeting. https://osf.io/9z538/

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31*(10), 1369–1385. https://doi.org/10.1177/0146167205275613

Hofmann, W., & Wilson, T. D. (2010). Consciousness, introspection, and the adaptive unconscious. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition. Measurement, theory, and applications* (pp. 197–215). New York, The Guilford Press.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960. https://doi.org/10.2307/2289064

Holland, P. W. (1988). [Employment discrimination and statistical science]: Comment: Causal mechanism or causal effect: Which is best for statistical science? *Statistical Science, 3*(2), 186–188. https://doi.org/10.1214/ss/1177012901

Holland, P. W. (2003). Causation and race. *ETS Research Report Series, 2003*(1), i–21. https://doi.org/10.1002/j.2333-8504.2003.tb01895.x

Holzer, H., & Ludwig, J. (2003). Measuring discrimination in education: Are methodologies from labor and markets useful? *The Teachers College Record, 105*(6), 1147–1178.

Homans, G. C. (1967). *The nature of social science*. New York, Harcourt, Brace & World.

Homans, G. C. (1970). The relevance of psychology to the explanation of social phenomena. In R. Borger & F. Cioffi (Eds.), *Explanation in the behavioural sciences* (pp. 313–329). Cambridge, Cambridge University Press.

Hout, M. (2012). Social and economic returns to college education in the United States. *Annual Review of Sociology, 38*, 379–400. https://doi.org/10.1146/annurev.soc.012809.102503

Hraba, J., Hagendoorn, L., & Hagendoorn, R. (1989). The ethnic hierarchy in the Netherlands: Social distance and social representation. *British Journal of Social Psychology, 28*(1), 57–69. https://doi.org/10.1111/j.2044-8309.1989.tb00846.x

Hume, D. (1738). *A treatise of human nature* (J. Bennett, Ed.). London.

Hummell, H. J., & Opp, K.-D. (1968). Sociology without sociology. *Inquiry, 11*(1), 205–226. https://doi.org/10.1080/00201746808601527

Hunkler, C. (2014). *Ethnische Ungleichheit beim Zugang zu Ausbildungsplätzen im dualen System*. Wiesbaden, Springer VS. https://doi.org/10.1007/978-3-658-05494-6

Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(1), 5–51. https://doi.org/10.1111/j.1467-985X.2012.01032.x/full

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. New York, Cambridge University Press.

International Labour Organization. (1990). *International standard classification of occupations: ISCO-88*. ILO. Geneva.

Ishida, H., Müller, W., & Ridge, J. M. (1995). Class origin, class destination, and education: A cross-national study of ten industrial nations. *American Journal of Sociology, 101*(1), 145–193. https://doi.org/10.1086/230701

Jackson, M. (2009). Disadvantaged through discrimination? The role of employers in social stratification. *The British Journal of Sociology, 60*(4), 669–692. https://doi.org/10.1111/j.1468-4446.2009.01270.x

Jackson, M., Goldthorpe, J. H., & Mills, C. (2005). Education, employers and class mobility. *Research in Social Stratification and Mobility, 23*, 3–33. https://doi.org/10.1016/S0276-5624(05)23001-9

Jacquemet, N., & Yannelis, C. (2012). Indiscriminate discrimination: A correspondence test for ethnic homophily in the Chicago labor market. *Labour Economics*, *19*(6), 824–832. https://doi.org/10.1016/j.labeco.2012.08.004

Jæger, M. M. (2011). "A thing of beauty is a joy forever"? Returns to physical attractiveness over the life course. *Social Forces*, *89*(3), 983–1003. https://doi.org/10.1093/sf/89.3.983

Jones, C. P. (2000). Levels of racism: A theoretic framework and a gardener's tale. *American Journal of Public Health*, *90*(8), 1212–1215.

Jones, J. M. (1972). *Prejudice and racism* (1st ed.). Reading, MA, Addison-Wesley Publishing Company.

Jones, J. M. (1997). *Prejudice and racism* (2nd ed.). New York, McGraw-Hill.

Judd, C. M., & Park, B. (1988). Out-group homogeneity: Judgments of variability at the individual and group levels. *Journal of Personality and Social Psychology*, *54*(5), 778–788. https://doi.org/10.1037/0022-3514.54.5.778

Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review*, *100*(1), 109–128. https://doi.org/10.1037/0033-295X.100.1.109

Judd, C. M., Ryan, C. S., & Park, B. (1991). Accuracy in the judgment of in-group and out-group variability. *Journal of Personality and Social Psychology*, *61*(3), 366.

Jussim, L. (1989). Teacher expectations: Self-fulfilling prophecies, perceptual biases, and accuracy. *Journal of Personality and Social Psychology*, *57*(3), 469–480.

Jussim, L. (2012). *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy* (1st ed.). New York, Oxford Univ Press.

Jussim, L., Cain, T. R., Crawford, J. T., Harber, K., & Cohen, F. (2009). The unbearable accuracy of stereotypes. In *Handbook of prejudice, stereotyping, and discrimination* (pp. 199–227). New York, NY, US, Psychology Press.

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (in)accuracy in perceptions of groups and individuals. *Current Directions in Psychological Science*, *24*(6), 490–497. https://doi.org/10.1177/0963721415605257

Jussim, L., Eccles, J., & Madon, S. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. *Advances in Experimental Social Psychology*, *28*, 281–388.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, *9*(2), 131–155. https://doi.org/10.1207/s15327957pspr0902_3

Jussim, L., McCauley, C. R., & Lee, Y.-T. (1995). Why study stereotype accuracy and inaccuracy? In Y.-T. Lee, L. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy. Toward appreciating group differences* (pp. 3–27). Washington, DC, American Psychological Association (APA).

Jussim, L., Robustelli, S. L., & Cain, T. R. (2009). Teacher expectations and self-fulfilling prophecies. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 349–380). New York & London, Routledge.

Kahneman, D., & Krueger, A. B. (2006). Developments in the measurement of subjective well-being. *The Journal of Economic Perspectives, 20*(1), 3–24. https://doi.org/10.1257/089533006776526030

Kalter, F. (2003). *Chancen, Fouls und Abseitsfallen. Migranten im deutschen Ligenfussball*. Wiesbaden, VS Verlag für Sozialwissenschaften.

Kalter, F. (2008). Ethnische Ungleichheit auf dem Arbeitsmarkt. In M. Abraham & T. Hinz (Eds.), *Arbeitsmarktsoziologie* (2nd ed., pp. 303–332). Wiesbaden, VS Verlag für Sozialwissenschaften.

Karing, C., Matthäi, J., & Artelt, C. (2011). Genauigkeit von Lehrerurteilen über die Lesekompetenz ihrer Schülerinnen und Schüler in der Sekundarstufe I – Eine Frage der Spezifität? *Zeitschrift Für Pädagogische Psychologie, 25*(3), 159–172. https://doi.org/10.1024/1010-0652/a000041

Karlson, K. B., Holm, A., & Breen, R. (2012). Comparing regression coefficients between same-sample nested models using logit and probit a new method. *Sociological Methodology, 42*(1), 286–313. https://doi.org/10.1177/0081175012444861

Katz, D., & Braly, K. (1933). Racial stereotypes of one hundred college students. *The Journal of Abnormal and Social Psychology, 28*(3), 280–290. https://doi.org/10.1037/h0074049

Katz, D., & Braly, K. (1935). Racial prejudice and racial stereotypes. *The Journal of Abnormal and Social Psychology, 30*(2), 175–193. https://doi.org/10.1037/h0059800

Kiss, D. (2013). Are immigrants and girls graded worse? Results of a matching approach. *Education Economics, 21*(5), 447–463. https://doi.org/10.1080/09645292.2011.585019

Klein, M. (2011). Trends in the association between educational attainment and class destinations in West Germany: Looking inside the service class. *Research in Social Stratification and Mobility, 29*(4), 427–444. https://doi.org/10.1016/j.rssm.2011.03.002

Kleinert, C. (2004). *Fremdenfeindlichkeit: Einstellungen junger Deutscher zu Migranten*. Wiesbaden, VS Verlag für Sozialwissenschaften.

Knowles, J., Persico, N., & Todd, P. (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, *109*(1), 203–229. https://doi.org/10.1086/318603

Knowles, L. L., & Prewitt, K. (Eds.). (1969). *Institutional racism in America*. Englewood Cliffs, NJ, Prentice Hall.

Kogan, I. (2004). Last hired, first fired? The unemployment dynamics of male immigrants in Germany. *European Sociological Review, 20*(5), 445–461. https://doi.org/10.1093/esr/jch037

Kogan, I. (2007). A study of immigrants' employment careers in West Germany using the sequence analysis technique. *Social Science Research, 36*(2), 491–511. https://doi.org/10.1016/j.ssresearch.2006.03.004

Krieger, L. H. (1995). The content of our categories: A cognitive bias approach to discrimination and equal employment opportunity. *Stanford Law Review*, *47*(6), 1161–1248. https://doi.org/10.2307/1229191

Krieger, L. H., & Fiske, S. T. (2006). Behavioral realism in employment discrimination law: Implicit bias and disparate treatment. *California Law Review, 94*(4), 997–1062. https://doi.org/10.2307/20439058

Kristen, C. (2002). Hauptschule, Realschule oder Gymnasium? *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 54*(3), 534–552. https://doi.org/10.1007/s11577-002-0073-2

Kristen, C. (2003). *School choice and ethnic school segregation: Primary school selection in Germany*. Waxmann Verlag.

Kristen, C. (2006a). *Ethnische Diskriminierung im deutschen Schulsystem? Theoretische Überlegungen und empirische Ergebnisse* (Discussion Paper SP IV 2006-601). Wissenschaftszentrum Berlin für Sozialforschung, Arbeitsstelle Interkulturelle Konflikte und Gesellschaftliche Integration. Berlin.

Kristen, C. (2006b). Ethnische Diskriminierung in der Grundschule? Die Vergabe von Noten und Bildungsempfehlungen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 58*(1), 79–97. https://doi.org/10.1007/s11575-006-0004-y

Kristen, C., & Dollmann, J. (2009). Sekundäre Effekte der ethnischen Herkunft: Kinder aus türkischen Familien am ersten Bildungsübergang. In J. Baumert, K. Maaz, & U. Trautwein (Eds.), *Bildungsentscheidungen* (pp. 205–229). Wiesbaden, VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92216-4_9

Kristen, C., Edele, A., Kalter, F., Kogan, I., Schulz, B., Stanat, P., & Will, G. (2011). 8 The education of migrants and their children across the life course (H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice, Eds.). *Zeitschrift für Erziehungswissenschaft, 14*(2), 121–137. https://doi.org/10.1007/s11618-011-0194-3

Kristen, C., & Granato, N. (2007). The educational attainment of the second generation in Germany social origins and ethnic inequality. *Ethnicities, 7*(3), 343–366.

Kroneberg, C. (2010). *Die Erklärung sozialen Handelns: Grundlagen und Anwendung einer integrativen Theorie*. Wiesbaden, VS Verlag für Sozialwissenschaften.

Kroneberg, C., & Kalter, F. (2012). Rational choice theory and empirical research: Methodological and theoretical contributions in Europe. *Annual Review of Sociology, 38*, 73–92. https://doi.org/10.1146/annurev-soc-071811-145441

Kroneberg, C., Yaish, M., & Stocké, V. (2010). Norms and rationality in electoral participation and in the rescue of jews in WWII. an application of the model of frame selection. *Rationality and Society, 22*(1), 3–36. https://doi.org/10.1177/1043463109355494

Lakatos, I. (1980). *The methodology of scientific research programmes* (J. Worall & G. Currie, Eds.; Philosophical Papers, Vol. 1). Cambridge University Press.

LaPiere, R. T. (1934). Attitudes vs. actions. *Social Forces, 13*(2), 230–237. https://doi.org/10.2307/2570339

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics, 92*(10), 2083–2105. https://doi.org/10.1016/j.jpubeco.2008.02.009

Lee, Y.-T., Jussim, L., & McCauley, C. R. (Eds.). (1995). *Stereotype accuracy. Toward appreciating group differences*. Washington, DC, American Psychological Association (APA).

Lepore, L., & Brown, R. (1997). Category and stereotype activation: Is prejudice inevitable? *Journal of Personality and Social Psychology, 72*, 275–287.

Levin, J., & Levin, W. (1982). *The functions of discrimination and prejudice* (2nd ed.). New York, Harper & Row Publishers.

Levitt, S. D. (2004). Testing theories of discrimination: Evidence from weakest link. *The Journal of Law and Economics, 47*(2), 431–452. https://doi.org/10.1086/425591

Leyens, J.-P., Schadron, G., & Yzerbyt, V. (1994). *Stereotypes and social cognition*. London; Thousand Oaks, Calif, SAGE Publications Ltd.

Lindahl, E. (2016). Are teacher assessments biased? Evidence from Sweden. *Education Economics, 24*(2), 224–238. https://doi.org/10.1080/09645292.2015.1014882

Linton, R. (1936). *The study of man. An introduction*. New York, Appleton Century Crofts, Inc.

Lippmann, W. (1922). *Public opinion*. Oxford, Harcourt Brace.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, SAGE.

Long, J. S. (2009). *Group comparisons in logit and probit using predicted probabilities.* Retrieved October 11, 2009, from http://www.indiana.edu/~jslsoc/files_research/groupdif/groupwithprobabilities/groups-with-prob-2009-06-25.pdf

López Bóo, F., Rossi, M. A., & Urzúa, S. S. (2013). The labor market return to an attractive face: Evidence from a field experiment. *Economics Letters, 118*(1), 170–172. https://doi.org/10.1016/j.econlet.2012.10.016

Lorenz, G., Gentrup, S., Kristen, C., Stanat, P., & Kogan, I. (2016). Stereotype bei Lehrkräften? Eine Untersuchung systematisch verzerrter Lehrererwartungen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 68*(1), 89–111. https://doi.org/10.1007/s11577-015-0352-3

Lorenzi-Cioldi, F., Eagly, A. H., & Stewart, T. L. (1995). Homogeneity of gender groups in memory. *Journal of Experimental Social Psychology, 31*(3), 193–217. https://doi.org/10.1006/jesp.1995.1010

Lott, B. (2002). Cognitive and behavioral distancing from the poor. *American Psychologist, 57*(2), 100–110. https://doi.org/10.1037/0003-066X.57.2.100

Lucas, S. R. (2008). *Theorizing discrimination in an era of contested prejudice. Discrimination in the United States* (Vol. 1). Philadelphia, Temple University Press.

Lüdemann, E., & Schwerdt, G. (2013). Migration background and educational tracking. *Journal of Population Economics, 26*(2), 455–481. https://doi.org/10.1007/s00148-012-0414-z

Maaz, K., Baeriswyl, F., & Trautwein, U. (2011). *Herkunft zensiert? Leistungsdiagnostik und soziale Ungleichheiten in der Schule. Eine Studie im Auftrag der Vodafone Stiftung Deutschland*. Vodafone Stiftung Deutschland. Düsseldorf. https://www.vodafone-stiftung.de/uploads/tx_newsjson/herkunft_zensiert_2012.pdf

Maaz, K., Baumert, J., Gresch, C., & McElvany, N. (Eds.). (2010). *Der Übergang von der Grundschule in die weiterführende Schule. Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten* (Bildungsforschung, Vol. 34). Bonn; Berlin, BMBF.

Maaz, K., Neumann, M., Trautwein, U., Wendt, W., Lehmann, R., & Baumert, J. (2008). Der Übergang von der Grundschule in die weiterführende Schule: Die Rolle von Schüler- und Klassenmerkmalen beim Einschätzen der individuellen Lernkompetenz durch die Lehrkräfte. *Schweizerische Zeitschrift Für Bildungswissenschaften, 30*(3), 519–548.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology, 51*, 93–120. https://doi.org/10.1146/annurev.psych.51.1.93

Marks, G. N. (2005). Accounting for immigrant non-immigrant differences in reading and mathematics in twenty countries. *Ethnic and Racial Studies*, *28*(5), 925–946. https://doi.org/10.1080/01419870500158943

Marsh, A., Sahin-Dikmen, M., & The European Opinion Research Group. (2003). *Discrimination in Europe* (Executive Summary No. 57.0). The European Commission.

Mayer, A. K. (2011). Does education increase political participation? *The Journal of Politics*, *73*(3), 633–645. https://doi.org/10.1017/S002238161100034X

McCauley, C., & Stitt, C. (1978). An individual and quantitative measure of stereotypes. *Journal of Personality*, *36*(9), 929–940. https://doi.org/10.1037/0022-3514.36.9.929

McClelland, D. C. (1961). *The achieving society*. New York, NY, Free Press.

McConahay, J. B. (1983). Modern racism and modern discrimination the effects of race, racial attitudes, and context on simulated hiring decisions. *Personality and Social Psychology Bulletin*, *9*(4), 551–558. https://doi.org/10.1177/0146167283094004

McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–125). San Diego, CA, US, Academic Press.

McCrudden, C. (1998). Merit principles. *Oxford Journal of Legal Studies*, *18*(4), 543–579. https://doi.org/10.1093/ojls/18.4.543

Meertens, R. W., & Pettigrew, T. F. (1997). Is subtle prejudice really prejudice? *The Public Opinion Quarterly*, *61*(1), 54–71. https://doi.org/10.2307/2749511

Meier, K. J., Stewart, J., & England, R. E. (1989). *Race, class, and education: The politics of second-generation discrimination*. Univ of Wisconsin Press.

Menger, C. (1883). *Problems of economics and sociology*. Urbana, University of Illinois Press.

Merton, R. K. (1949). Discrimination and the American creed. In R. M. McIver (Ed.), *Discrimination and national welfare* (pp. 99–126). New York & London, Harper & Brothers.

Meyer, A. (2015). Does education increase pro-environmental behavior? Evidence from Europe. *Ecological Economics*, *116*, 108–121. https://doi.org/10.1016/j.ecolecon.2015.04.018

Mickelson, R. (2003). When are racial disparities in education the result of racial discrimination? A social science perspective. *The Teachers College Record*, *105*(6), 1052–1086.

Mill, J. S. (1843). *A system of logic. Ratiocinative and inductive. Collected works* (Vol. VII-VIII). Toronto, Toronto University Press.

Miller, K., Chepp, V., Willson, S., & Padilla, J. L. (2014). *Cognitive interviewing methodology*. Hoboken, New Jersey, John Wiley & Sons.

Mincer, J. (1991, September). *Education and unemployment* (Working Paper No. 3838). National Bureau of Economic Research. https://doi.org/10.3386/w3838

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review, 26*(1), 67–82. https://doi.org/10.1093/esr/jcp006

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge University Press.

Mücke, S., & Schründer-Lenzen, A. (2008). Zur Parallelität der Schulleistungsentwicklung von Jungen und Mädchen im Verlauf der Grundschule. In B. Rendtorff & A. Prengel (Eds.), *Kinder und ihr Geschlecht* (pp. 135–146). Frankfurt am Main, Verlag Barbara Budrich, Opladen.

Müller, W., & Pollak, R. (2004). Social mobility in west Germany: The long arms of history discovered? In R. Breen (Ed.), *Social mobility in Europe* (pp. 77–114). Oxford, Oxford University Press.

Myrdal, G. (1944). *An American dilemma: The negro problem and modern democracy*. New Brunswick, NJ, Transaction Publishers.

Nelson, T. D. (2006). *The psychology of prejudice* (2nd ed.). Boston, Pearson.

Neumark, D. (2012). Detecting discrimination in audit and correspondence studies. *Journal of Human Resources, 47*(4), 1128–1157. https://doi.org/10.3368/jhr.47.4.1128

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *The Journal of Applied Psychology, 93*(6), 1314–1334. https://doi.org/10.1037/a0012702

Nier, J. A., & Gaertner, S. L. (2012). The challenge of detecting contemporary forms of discrimination. *Journal of Social Issues, 68*(2), 207–220.

North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge; New York, Cambridge University Press.

Nosek, B. A. (2007). Implicit–explicit relations. *Current Directions in Psychological Science, 16*(2), 65–69. https://doi.org/10.1111/j.1467-8721.2007.00477.x

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *Eu-*

ropean Review of Social Psychology, 18(1), 36–88. https://doi.org/10.1080/10463280701489053

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review, 14*(3), 693–709. https://doi.org/10.2307/2525981

OECD. (2010). *PISA 2009 results: Overcoming social background – equity in learning opportunities and outcomes (volume II)*. Paris, OECD Publishing. https://doi.org/10.1787/9789264091504-en

OECD. (2016a). *Education at a glance 2016*. Paris, OECD Publishing. https://doi.org/10.1787/eag-2016-en

OECD. (2016b). *PISA 2015 results (volume I): Excellence and equity in education*. Paris, OECD Publishing. https://doi.org/10.1787/9789264266490-en

Olczyk, M. (2016). Migranten und ihre Nachkommen im deutschen Bildungssystem: Ein aktueller Überblick. In C. Diehl, C. Hunkler, C. Kristen, J. Seuring, G. Will, & S. Zinn (Eds.), *Ethnische Ungleichheiten im Bildungsverlauf* (pp. 33–70). Wiesbaden, Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-04322-3

Oppenheimer, D. B. (1993). Negligent discrimination. *University of Pennsylvania Law Review, 141*(3), 899–972. https://doi.org/10.2307/3312446

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Oxford, England, Univer. Illinois Press.

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual Review of Sociology, 34,* 181–209. https://doi.org/10.1146/annurev.soc.33.040406.131740

Park, B., & Judd, C. M. (1990). Measures and models of perceived group variability. *Journal of Personality, 59*(2), 173–191. https://doi.org/10.1037/0022-3514.59.2.173

Park, B., & Rothbart, M. (1982). Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. *Journal of Personality, 42*(6), 1051–1068. https://doi.org/10.1037/0022-3514.42.6.1051

Park, R. E. (1924). The concept of social distance as applied to the study of racial attitudes and racial relations. *Journal of Applied Sociology, 8*, 339–344.

Parrillo, V. N., & Donoghue, C. (2005). Updating the bogardus social distance studies: A new national survey. *The Social Science Journal, 42*(2), 257–271. https://doi.org/10.1016/j.soscij.2005.03.011

Parrillo, V. N., & Donoghue, C. (2013). The national social distance study: Ten years later. *Sociological Forum, 28*(3), 597–614. https://doi.org/10.1111/socf.12039

Parsons, T. (1940). An analytical approach to the theory of social stratification. *American Journal of Sociology, 45*(6), 841–862. https://doi.org/10.1086/218489

Parsons, T. (1950). The prospects of sociological theory. *American Sociological Review, 15*(1), 3–16. https://doi.org/10.2307/2086393

Pearl, J. (2001). Direct and indirect effects, In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc. Retrieved March 24, 2020, from https://ftp.cs.ucla.edu/pub/stat_ser/R273-U.pdf

Pearl, J. (2003). Statistics and causal inference: A review. *Test, 12*(2), 281–345. https://doi.org/10.1007/BF02595718

Pearl, J. (2009). *Causality* (2nd revised edition). Cambridge, UK; New York, NY, Cambridge University Press.

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods, 19*(4), 459–481. https://doi.org/10.1037/a0036434

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. Chichester, UK ; Hoboken, NJ, John Wiley; Sons Ltd.

Peresie, J. L. (2009). Toward a coherent test for disparate impact discrimination. *Indiana Law Journal, 84*, 773.

Pettigrew, T. F., & Meertens, R. W. (1995). Subtle and blatant prejudice in western Europe. *European Journal of Social Psychology, 25*(1), 57–75. https://doi.org/10.1002/ejsp.2420250106

Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review, 62*(4), 659–661.

Pincus, F. L. (1996). Discrimination comes in many forms: Individual, institutional, and structural. *American Behavioral Scientist, 40*(2), 186–194. https://doi.org/10.1177/0002764296040002009

Popper, K. R. (1945). *The open society and its enemies* (5th revised ed). Princeton, Princeton University Press.

Popper, K. R. (1957). *The poverty of historicism*. London, Routledge & Kegan Paul.

Popper, K. R. (2004). *The logic of scientific discovery*. London, Routledge. (Original work published 1959)

Prenzel, M., Sälzer, C., Klieme, E., & Köller, O. (Eds.). (2013). *PISA 2012: Fortschritte und Herausforderungen in Deutschland*. Münster, Waxmann.

Psacharopoulos, G., & Patrinos, H. A. (2004). Returns to investment in education: A further update. *Education Economics, 12*(2), 111–134. https://doi.org/10.1080/0964529042000239140

Quillian, L. (1995). Prejudice as a response to perceived group threat: Population composition and anti-immigrant and racial prejudice in Europe. *American Sociological Review*, 586–611. https://doi.org/10.2307/2096296

Quillian, L. (2006). New approaches to understanding racial prejudice and discrimination. *Annual Review of Sociology*, *32*, 299–328. https://doi.org/10.1146/annurev.soc.32.061604.123132

Quillian, L. (2008). Does unconscious racism exist? *Social Psychology Quarterly*, *71*(1), 6–11. https://doi.org/10.1177/019027250807100103

Raub, W., Buskens, V., & Van Assen, M. A. L. M. (2011). Micro-macro links and microfoundations in sociology. *The Journal of Mathematical Sociology*, *35*(1), 1–25. https://doi.org/10.1080/0022250X.2010.532263

Rawls, J. (1971). *A theory of justice* (1st ed.). Cambridge, Mass, Harvard University Press.

Ready, D. D., & Wright, D. L. (2010). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal*, *48*(2), 335–360. https://doi.org/10.3102/0002831210374874

Reiss, K., Sälzer, C., Schiepe-Tiska, A., Klieme, E., & Köller, O. (Eds.). (2016). *PISA 2015: Eine Studie zwischen Kontinuität und Innovation*. Münster; New York, Waxmann.

Reskin, B. F. (2003). Including mechanisms in our models of ascriptive inequality: 2002 presidential address. *American Sociological Review*, *68*(1), 1–21. https://doi.org/10.2307/3088900

Reskin, B. F. (2012). The race discrimination system. *Annual Review of Sociology*, *38*(1), 17–35. https://doi.org/10.1146/annurev-soc-071811-145508

Rindermann, H. (2007). The g-factor of international cognitive ability comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality*, *21*(5), 667–706. https://doi.org/10.1002/per.634

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, *3*(2), 143–155. https://doi.org/10.1097/00001648-199203000-00013

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (Eds.). (1999). *Measures of political attitudes*. San Diego, Academic Press.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/2515245917745629

Rubin, D. B. (1975). Bayesian inference for causality: The importance of randomization. *Proceedings of the Social Statistics Section of the American Statistical Association*, 233–239.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*(396), 961–962. https://doi.org/10.2307/2289065

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics, 31*(2), 161–170. https://doi.org/10.1111/j.1467-9469.2004.02-123.x

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics, 2*(3), 808–840. https://doi.org/10.1214/08-AOAS187

Rubin, M., & Hewstone, M. (1998). Social identity theory's self-esteem hypothesis: A review and some suggestions for clarification. *Personality and Social Psychology Review, 2*(1), 40–62. https://doi.org/10.1207/s15327957pspr0201_3

Rubovitz, P. C., & Maehr, M. L. (1973). Pygmalion black and white. *Journal of Personality and Social Psychology, 25*(2), 210–218. https://doi.org/10.1037/h0034080

Rutherglen, G. (1987). Disparate impact under title VII: An objective theory of discrimination. *Virginia Law Review, 73*(7), 1297–1345. https://doi.org/10.2307/1072940

Ryan, C. (2002). Stereotype accuracy. *European Review of Social Psychology, 13*(1), 75–109. https://doi.org/10.1080/10463280240000037

Ryan, C. S., & Bogart, L. M. (2001). Longitudinal changes in the accuracy of new group members' in-group and out-group stereotypes. *Journal of Experimental Social Psychology, 37*(2), 118–133. https://doi.org/10.1006/jesp.2000.1439

Schaeffer, M., Höhne, J., & Teney, C. (2016). Income advantages of poorly qualified immigrant minorities: Why school dropouts of Turkish origin earn more in Germany. *European Sociological Review, 32*(1), 93–107. https://doi.org/10.1093/esr/jcv091

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review, 115*(2), 336.

Schneider, D. J. (2004). *The psychology of stereotyping* (1st ed.). New York & London, The Guilford Press.

Schneider, S. L. (2008). Anti-immigrant attitudes in Europe: Outgroup size and perceived ethnic threat. *European Sociological Review, 24*(1), 53–67. https://doi.org/10.1093/esr/jcm034

Schneider, T. (2011). Die Bedeutung der sozialen Herkunft und des Migrationshintergrundes für Lehrerurteile am Beispiel der Grundschulempfehlung. *Zeitschrift*

Für Erziehungswissenschaft, *14*(3), 371–396. https://doi.org/10.1007/s11618-011-0221-4

Schnepf, S. V. (2007). Immigrants' educational disadvantage: An examination across ten countries and three surveys. *Journal of Population Economics*, *20*(3), 527–545. https://doi.org/10.1007/s00148-006-0102-y

Schulze, A., & Schiener, J. (2011). Lehrerurteile und Bildungsgerechtigkeit. *Zeitschrift für Soziologie der Erziehung und Sozialisation*, *31*(2).

Schütz, H., & Six, B. (1996). How strong is the relationship between prejudice and discrimination? A meta-analytic answer. *International Journal of Intercultural Relations*, *20*(3), 441–462. https://doi.org/10.1016/0147-1767(96)00028-4

Schwarz, N., Strack, F., & Mai, H.-P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly*, *55*(1), 3–23. https://doi.org/10.1086/269239

Schwarz, N., & Sudman, S. (1996). *Answering questions: Methodology for determining cognitive and communicative processes in survey research*. San Francisco, Jossey Bass.

Sears, D. O., & Henry, P. (2005). Over thirty years later: A contemporary look at symbolic racism. In *Advances in experimental social psychology* (pp. 95–150). Academic Press.

Sen, A. (1999). Merit and justice. In K. J. Arrow, S. Bowles, & S. Durlauf (Eds.), *Meritocracy and economic inequality*. Princeton, Princeton University Press.

Sen, M., & Wasow, O. (2016). Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, *19*(1), 499–522. https://doi.org/10.1146/annurev-polisci-032015-010015

Sewell, W. H., Haller, A. O., & Ohlendorf, G. W. (1970). The educational and early occupational status attainment process: Replication and revision. *American Sociological Review*, *35*(6), 1014–1027. https://doi.org/10.2307/2093379

Sewell, W. H., Haller, A. O., & Portes, A. (1969). The educational and early occupational attainment process. *American Sociological Review*, *34*(1), 82–92. https://doi.org/10.2307/2092789

Sewell, W. H., & Hauser, R. M. (1975). *Education, occupation and earnings: Achievement in the early career*. New York, Academic Press Inc.

Shavit, Y., & Blossfeld, H.-P. (1993). *Persistent inequality: Changing educational attainment in thirteen countries*. Boulder, Colo., Westview Press.

Sidanius, J., & Pratto, F. (1999, July 28). *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge, UK; New York, NY, Cambridge University Press.

Simpson, G. E., & Yinger, J. M. (1972). *Racial and cultural minorities: An analysis of prejudice and discrimination* (4th). New York, NY, Harper & Row Publishers.

Smith, T. W., & Dempsey, G. R. (1983). The polls: Ethnic social distance and prejudice. *The Public Opinion Quarterly, 47*(4), 584–600. https://doi.org/10.2307/2748673

Snyder, M., & Swann, W. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality, 36*(11), 1202–1212. https://doi.org/10.1037/0022-3514.36.11.1202

Sobel, M. E. (1998). Causal inference in statistical models of the process of socioeconomic achievement a case study. *Sociological Methods & Research, 27*(2), 318–348. https://doi.org/10.1177/0049124198027002006

Sprietsma, M. (2013). Discrimination in grading: Experimental evidence from primary school teachers. *Empirical Economics, 45*(1), 523–538. https://doi.org/10.1007/s00181-012-0609-x

Stahl, N. (2007). Schülerwahrnehmung und -beurteilung durch Lehkräfte. In H. Ditton (Ed.), *Kompetenzaufbau und Laufbahnen im Schulsystem* (pp. 171–198). Waxmann Verlag.

Stanat, P., Pant, H. A., Böhme, K., & Richter, D. (Eds.). (2012, October 5). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik: Ergebnisse des IQB-Ländervergleichs 2011*. Waxmann.

Stangor, C., Sullivan, L. A., & Ford, T. E. (1991). Affective and cognitive determinants of prejudice. *Social Cognition, 9*(4), 359–380. https://doi.org/10.1521/soco.1991.9.4.359

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797–811. https://doi.org/10.1037/0022-3514.69.5.797

Steinbach, A. (2004). *Soziale Distanz: Ethnische Grenzziehung und die Eingliederung von Zuwanderern in Deutschland*. Springer-Verlag.

Stocké, V., Blossfeld, H.-P., Hoenig, K., & Sixt, M. (2011). 7 Social inequality and educational decisions in the life course (H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice, Eds.). *Zeitschrift Für Erziehungswissenschaft, 14*(2), 103–119. https://doi.org/10.1007/s11618-011-0193-4

Storm, I., Sobolewska, M., & Ford, R. (2017). Is ethnic prejudice declining in Britain? Change in social distance attitudes among ethnic majority and minority Britons. *The British Journal of Sociology, 68*(3), 410–434. https://doi.org/10.1111/1468-4446.12250

Stricker, L. J., Rock, D. A., & Bridgeman, B. (2015). Stereotype threat, inquiring about test takers' race and gender, and performance on low-stakes tests in a large-

scale assessment. *ETS Research Report Series, 2015*(1), 1–12. https://doi.org/10.1002/ets2.12046

Südkamp, A., Kaiser, J., & Moller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology August 2012, 104*(3), 743–762. https://doi.org/10.1037/a0027627

Sumner, W. G. (1906). *Folkways: A study of mores, manners, customs, and morals*. New York, Ginn.

Sundstrom, W. A. (1990). Half a career: Discrimination & railroad internal labor markets. *Industrial Relations: A Journal of Economy and Society, 29*(3), 423–440. https://doi.org/10.1111/j.1468-232X.1990.tb00762.x

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American, 223*(5), 96–102.

Tajfel, H. (1982). *Social identity and intergroup relations*. Cambridge, Cambridge University Press.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations*. Chicago, IL, Nelson-Hall Publishers.

Talaska, C. A., Fiske, S. T., & Chaiken, S. (2008). Legitimating racial discrimination: Emotions, not beliefs, best predict discrimination in a meta-analysis. *Social Justice Research, 21*(3), 263–296. https://doi.org/10.1007/s11211-008-0071-2

Taylor, M. C. (1979). Race, sex, and the expression of self-fulfilling prophecies in a laboratory teaching situation. *Journal of Personality and Social Psychology, 37*(6), 897–912. https://doi.org/10.1037/0022-3514.37.6.897

Tobisch, A., & Dresel, M. (2017). Negatively or positively biased? dependencies of teachers' judgments and expectations based on students' ethnic and social backgrounds. *Social Psychology of Education, 20*, 731–752. https://doi.org/10.1007/s11218-017-9392-z

Tobisch, A., & Dresel, M. (2020). Correction to: Negatively or positively biased? dependencies of teachers' judgments and expectations based on students' ethnic and social backgrounds. *Social Psychology of Education*. https://doi.org/10.1007/s11218-020-09548-0

Triandis, H. C., & Triandis, L. M. (1960). Race, social class, religion, and nationality as determinants of social distance. *The Journal of Abnormal and Social Psychology, 61*(1), 110–118. https://doi.org/10.1037/h0041734

Triandis, H. C., & Triandis, L. M. (1962). A cross-cultural study of social distance. *Psychological Monographs: General and Applied, 76*(21), 1–21. https://doi.org/10.1037/h0093836

Troyna, B., & Williams, J. (2012). *Racism, education and the state*. Abingdon, Routledge.

Udehn, L. (2002). The changing face of methodological individualism. *Annual Review of Sociology, 28*, 479–507. https://doi.org/10.1146/annurev.soc.28.110601.140938

Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria redefining merit to justify discrimination. *Psychological Science, 16*(6), 474–480. https://doi.org/10.1111/j.0956-7976.2005.01559.x

VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. New York, Oxford University Press.

VanderWeele, T. J., & Hernán, M. A. (2012). Causal effects and natural laws: Towards a conceptualization of causal counterfactuals for nonmanipulable exposures, with application to the effects of race and sex. In C. Berzuini, P. Dawid, & L. Bernardinelli (Eds.), *Causality* (pp. 101–113). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119945710.ch9/summary

van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of Education Review, 30*(5), 1045–1058. https://doi.org/10.1016/j.econedurev.2011.05.008

Vargas, P. T., Sekaquaptewa, D., & von Hippel, W. (2007). Armed only with paper and pencil: "Low-tech"implicit measures of attitudes, prejudice, and self-esteem. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 103–124). New York, NY, The Guilford Press.

Verba, S., Schlozman, K. L., & Brady, H. E. (1995). *Voice and equality: Civic voluntarism in American politics*. Cambridge, MA, Harvard University Press.

Vinacke, W. E. (1957). Stereotypes as social concepts. *The Journal of Social Psychology, 46*(2), 229–243. https://doi.org/10.1080/00224545.1957.9714322

von dem Knesebeck, O., Verde, P. E., & Dragano, N. (2006). Education and health in 22 European countries. *Social Science & Medicine, 63*(5), 1344–1351. https://doi.org/10.1016/j.socscimed.2006.03.043

Wagner, U., Dick, R. v., Petzel, T., & Auernheimer, G. (2001). Der Umgang von Lehrerinnen und Lehrern mit interkulturellen Konflikten. In P. D. G. Auernheimer, D. R. v. Dick, D.-P. T. Petzel, & P. D. U. Wagner (Eds.), *Interkulturalität im Arbeitsfeld Schule* (pp. 17–40). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-80867-7_2

Wagner, U., & Zick, A. (1995). The relation of formal education to ethnic prejudice: Its reliability, validity and explanation. *European Journal of Social Psychology, 25*(1), 41–56. https://doi.org/10.1002/ejsp.2420250105/full

Walter, O. (2009). Herkunftsassoziierte Disparitäten im Lesen, der Mathematik und den Naturwissenschaften: ein Vergleich zwischen PISA 2000, PISA 2003 und PISA 2006. In M. Prenzel & J. Baumert (Eds.), *Vertiefende Analysen zu PISA 2006* (pp. 149–168). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91815-0_8

Wang, X., & Sobel, M. E. (2013). New perspectives on causal mediation analysis. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 215–242). Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3_12

Weber, M. (1922). *Gesammelte Aufsätze zur Wissenschaftslehre*. Tübingen, Mohr. https://www.neps-data.de/Portals/0/Survey%20Papers/SP_III.pdf

Weichselbaumer, D. (2016). Discrimination against female migrants wearing headscarves. *IZA Discussion Paper Series*, (10217).

Weijters, B., Geuens, M., & Schillewaert, N. (2009). The proximity effect: The role of inter-item distance on reverse-item bias. *International Journal of Research in Marketing*, *26*(1), 2–12. https://doi.org/10.1016/j.ijresmar.2008.09.003

Wendt, H., Bos, W., Köller, O., Schwippert, K., & Kasper, D. (Eds.). (2016). *TIMSS 2015: Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (1st ed.). Münster; New York, Waxmann.

Wenz, S. E. (2009). *Discrimination in the German education system: Teachers' track recommendations at the end of primary school and educational opportunity* (Diploma thesis). University of Mannheim.

Wenz, S. E., & Hoenig, K. (2020). Ethnic and social class discrimination in education: Experimental evidence from germany. *Research in Social Stratification and Mobility*, *65*. https://doi.org/10.1016/j.rssm.2019.100461

Wenz, S. E., Olczyk, M., & Lorenz, G. (2016). Measuring teachers' stereotypes in the NEPS. *NEPS Survey Papers*, (3).

Whitley, B. E. J. (1999). Right-wing authoritarianism, social dominance orientation, and prejudice. *Journal of Personality*, *77*(1), 126–134.

Wight, C. (2003). The agent–structure problem and institutional racism. *Political Studies*, *51*(4), 706–721. https://doi.org/10.1111/j.0032-3217.2003.00454.x

Williams, J. (1985). Redefining institutional racism. *Ethnic and Racial Studies*, *8*(3), 323–348. https://doi.org/10.1080/01419870.1985.9993490

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA, SAGE Publications, Inc.

Wippler, R. (1978). The structural-individualistic approach in dutch sociology: Toward and explanatory social science. *The Netherlands Journal of Sociology*, *14*(2), 135–155.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology*, *72*(2), 262.

Wong, J. S., & Penner, A. M. (2016). Gender and the returns to attractiveness. *Research in Social Stratification and Mobility*, *44*, 113–123. https://doi.org/10.1016/j.rssm.2016.04.002

Wooldridge, J. (2013). *Introductory econometrics: A modern approach* (5th ed.). Cengage Learning.

Wyer, N. A., Sadler, M. S., & Judd, C. M. (2002). Contrast effects in stereotype formation and change: The role of comparative context. *Journal of Experimental Social Psychology*, *38*(5), 443–458. https://doi.org/10.1016/S0022-1031(02)00010-0

Yang, Y. (2008). Social inequalities in happiness in the United States, 1972 to 2004: An age-period-cohort analysis. *American Sociological Review*, *73*(2), 204–226. https://doi.org/10.1177/000312240807300202

Young, M. (1958). *The rise of the meritocracy 1870–2033. An essay on education and equality*. Harmondsworth, Penguin Books.

Zanna, M. P., & Rempel, J. K. (1988). Attitudes: A new look at an old concept. In D. Bar-Tal & A. W. Kruglanski (Eds.), *The social psychology of knowledge* (pp. 315–334). Cambridge, Cambridge University Press.

Zarate, M. A., & Smith, E. R. (1990). Person categorization and stereotyping. *Social Cognition*, *8*(2), 161–185. https://doi.org/10.1521/soco.1990.8.2.161

Zschirnt, E., & Ruedin, D. (2016). Ethnic discrimination in hiring decisions: A meta-analysis of correspondence tests 1990–2015. *Journal of Ethnic and Migration Studies*, *42*(7), 1115–1134. https://doi.org/10.1080/1369183X.2015.1133279

In *Discrimination in Education* Sebastian E. Wenz makes methodological, theoretical, and empirical contributions to the study of discriminatory behavior by teachers towards students of different ethnicity, social class background, and gender. Wenz reviews different motivations to study discrimination in education and beyond, provides an in-depth discussion of numerous definitions of discrimination, and shows the potentials and limitations of theories from different disciplines to explain discrimination on the individual level and inequality between groups in the education system. In the empirical part of the book, Wenz examines the two major determinants of discrimination – prejudice and stereotypes – using data from the German General Social Survey (GGSS/ALLBUS) and the National Educational Panel Study (NEPS). The major empirical contribution in *Discrimination in Education* is an experimental study with a random sample of over 200 elementary school teachers. In this study, Wenz addresses several shortcomings of prior research. After discussing the complex findings of his experiment, Wenz concludes by suggesting several routes future research should take to gather more evidence on discrimination in education, its determinants and consequences.

In *Discrimination in Education* leistet Sebastian E. Wenz methodische, theoretische und empirische Beiträge zur Untersuchung diskriminierenden Verhaltens von Lehrkräften gegenüber Schülerinnen und Schülern unterschiedlicher Ethnizität, sozialer Herkunft und Geschlecht. Wenz beschreibt verschiedene Motivationen zur Untersuchung von Diskriminierung innerhalb und außerhalb des Bildungswesens, diskutiert ausführlich zahlreiche Definitionen von Diskriminierung und zeigt die Möglichkeiten und Grenzen von Theorien aus verschiedenen Disziplinen zur Erklärung von Diskriminierung auf der individuellen Ebene und der Ungleichheit zwischen Gruppen im Bildungswesen auf. Im empirischen Teil des Buches untersucht Wenz die beiden wichtigsten Determinanten von Diskriminierung – Vorurteile und Stereotype – anhand von Daten der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) und dem Nationalen Bildungspanel (NEPS). Der wichtigste empirische Beitrag in *Discrimination in Education* ist eine experimentelle Studie mit einer Zufallsstichprobe von über 200 Grundschullehrkräften. Mit seinem Studiendesign adressiert Wenz verschiedene Limitationen der bisherigen Forschung. Nachdem er die vielschichtigen Ergebnisse seines Experiments diskutiert hat, macht Wenz abschließend verschiedene Vorschläge, wie zukünftige Studien weitere Evidenz zum Problemkomplex Diskriminierung im Bildungswesen sammeln können.

gesis
Leibniz-Institut
für Sozialwissenschaften

ISBN 978-3-86819-044-1 (print)
ISBN 978-3-86819-043-4 (eBook)
ISSN 1869-2869
15,- EUR